美国数学会经典影印系列



Markov Chains and Mixing Times

Markov 链与混合时间

David A. Levin
Yuval Peres
Elizabeth L. Wilmer

高等教育出版社

美国数学会经典影印系列



Markov Chains and Mixing Times

Markov 链与混合时间

David A. Levin
Yuval Peres
Elizabeth L. Wilmer

With a chapter on "Coupling from the Past" by
James G. Propp and David B. Wilson

高等教育出版社 · 北京

图字：01-2016-2517 号

Markov 链与混合时间
Markov Lian yu Hunhe Shijian

图书在版编目 (CIP) 数据

Markov 链与混合时间 = Markov Chains and Mixing Times : 英文 / (美) 大卫 · A · 莱文 (David A. Levin), (以) 尤瓦尔 · 佩雷斯 (Yuval Peres) , (美) 伊丽莎白 · L · 威尔默 (Elizabeth L. Wilmer) 著 . — 影印本 . — 北京 : 高等教育出版社 , 2017.4
ISBN 978-7-04-046994-3
Ⅰ. ①M… Ⅱ. ①大… ②尤… ③伊… Ⅲ. ①马尔柯夫链—英文Ⅳ. ①O211.62
中国版本图书馆 CIP 数据核字 (2016) 第 326796 号

策划编辑 李 鹏　　责任编辑 李 鹏
封面设计 张申申　　责任印制 赵义民

出版发行 高等教育出版社
社址 北京市西城区德外大街4号
邮政编码 100120
购书热线 010-58581118
咨询电话 400-810-0598
网址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn
印刷 北京中科印刷有限公司

开本 787mm × 1092 mm 1/16
印张 24.75
字数 620 千字
版次 2017 年 4 月第 1 版
印次 2017 年 4 月第 1 次印刷
定价 169.00元



[物 料 号 46994-00]

Preface

Markov first studied the stochastic processes that came to be named after him in 1906. Approximately a century later, there is an active and diverse interdisciplinary community of researchers using Markov chains in computer science, physics, statistics, bioinformatics, engineering, and many other areas.

The classical theory of Markov chains studied *fixed* chains, and the goal was to estimate the rate of convergence to stationarity of the distribution at time t, as $t \to \infty$. In the past two decades, as interest in chains with large state spaces has increased, a different asymptotic analysis has emerged. Some target distance to the stationary distribution is prescribed; the number of steps required to reach this target is called the *mixing time* of the chain. Now, the goal is to understand how the mixing time grows as the size of the state space increases.

The modern theory of Markov chain mixing is the result of the convergence, in the 1980's and 1990's, of several threads. (We mention only a few names here; see the chapter Notes for references.)

For statistical physicists Markov chains become useful in Monte Carlo simulation, especially for models on finite grids. The mixing time can determine the running time for simulation. However, Markov chains are used not only for simulation and sampling purposes, but also as models of dynamical processes. Deep connections were found between rapid mixing and spatial properties of spin systems, e.g., by Dobrushin, Shlosman, Stroock, Zegarlinski, Martinelli, and Olivieri.

In theoretical computer science, Markov chains play a key role in sampling and approximate counting algorithms. Often the goal was to prove that the mixing time is polynomial in the logarithm of the state space size. (In this book, we are generally interested in more precise asymptotics.)

At the same time, mathematicians including Aldous and Diaconis were intensively studying card shuffling and other random walks on groups. Both spectral methods and probabilistic techniques, such as coupling, played important roles. Alon and Milman, Jerrum and Sinclair, and Lawler and Sokal elucidated the connection between eigenvalues and expansion properties. Ingenious constructions of "expander" graphs (on which random walks mix especially fast) were found using probability, representation theory, and number theory.

In the 1990's there was substantial interaction between these communities, as computer scientists studied spin systems and as ideas from physics were used for sampling combinatorial structures. Using the geometry of the underlying graph to find (or exclude) bottlenecks played a key role in many results.

There are many methods for determining the asymptotics of convergence to stationarity as a function of the state space size and geometry. We hope to present these exciting developments in an accessible way.

We will only give a taste of the applications to computer science and statistical physics; our focus will be on the common underlying mathematics. The prerequisites are all at the undergraduate level. We will draw primarily on probability and linear algebra, but we will also use the theory of groups and tools from analysis when appropriate.

Why should mathematicians study Markov chain convergence? First of all, it is a lively and central part of modern probability theory. But there are ties to several other mathematical areas as well. The behavior of the random walk on a graph reveals features of the graph's geometry. Many phenomena that can be observed in the setting of finite graphs also occur in differential geometry. Indeed, the two fields enjoy active cross-fertilization, with ideas in each playing useful roles in the other. Reversible finite Markov chains can be viewed as resistor networks; the resulting discrete potential theory has strong connections with classical potential theory. It is amusing to interpret random walks on the symmetric group as card shuffles—and real shuffles have inspired some extremely serious mathematics—but these chains are closely tied to core areas in algebraic combinatorics and representation theory.

In the spring of 2005, mixing times of finite Markov chains were a major theme of the multidisciplinary research program *Probability, Algorithms, and Statistical Physics,* held at the Mathematical Sciences Research Institute. We began work on this book there.

Overview

We have divided the book into two parts.

In **Part I**, the focus is on techniques, and the examples are illustrative and accessible. Chapter 1 defines Markov chains and develops the conditions necessary for the existence of a unique stationary distribution. Chapters 2 and 3 both cover examples. In Chapter 2, they are either classical or useful—and generally both; we include accounts of several chains, such as the gambler's ruin and the coupon collector, that come up throughout probability. In Chapter 3, we discuss Glauber dynamics and the Metropolis algorithm in the context of "spin systems." These chains are important in statistical mechanics and theoretical computer science.

Chapter 4 proves that, under mild conditions, Markov chains do, in fact, converge to their stationary distributions and defines *total variation distance* and *mixing time*, the key tools for quantifying that convergence. The techniques of Chapters 5, 6, and 7, on coupling, strong stationary times, and methods for lower bounding distance from stationarity, respectively, are central to the area.

In Chapter 8, we pause to examine card shuffling chains. Random walks on the symmetric group are an important mathematical area in their own right, but we hope that readers will appreciate a rich class of examples appearing at this stage in the exposition.

Chapter 9 describes the relationship between random walks on graphs and electrical networks, while Chapters 10 and 11 discuss hitting times and cover times.

Chapter 12 introduces eigenvalue techniques and discusses the role of the relaxation time (the reciprocal of the spectral gap) in the mixing of the chain.

In **Part II**, we cover more sophisticated techniques and present several detailed case studies of particular families of chains. Much of this material appears here for the first time in textbook form.

Chapter 13 covers advanced spectral techniques, including comparison of Dirichlet forms and Wilson's method for lower bounding mixing.

Chapters 14 and 15 cover some of the most important families of "large" chains studied in computer science and statistical mechanics and some of the most important methods used in their analysis. Chapter 14 introduces the path coupling method, which is useful in both sampling and approximate counting. Chapter 15 looks at the Ising model on several different graphs, both above and below the critical temperature.

Chapter 16 revisits shuffling, looking at two examples—one with an application to genomics—whose analysis requires the spectral techniques of Chapter 13.

Chapter 17 begins with a brief introduction to martingales and then presents some applications of the evolving sets process.

Chapter 18 considers the cutoff phenomenon. For many families of chains where we can prove sharp upper and lower bounds on mixing time, the distance from stationarity drops from near 1 to near 0 over an interval asymptotically smaller than the mixing time. Understanding why cutoff is so common for families of interest is a central question.

Chapter 19, on lamplighter chains, brings together methods presented throughout the book. There are many bounds relating parameters of lamplighter chains to parameters of the original chain: for example, the mixing time of a lamplighter chain is of the same order as the cover time of the base chain.

Chapters 20 and 21 introduce two well-studied variants on finite discrete time Markov chains: continuous time chains and chains with countable state spaces. In both cases we draw connections with aspects of the mixing behavior of finite discrete-time Markov chains.

Chapter 22, written by Propp and Wilson, describes the remarkable construction of coupling from the past, which can provide exact samples from the stationary distribution.

Chapter 23 closes the book with a list of open problems connected to material covered in the book.

For the Reader

Starred sections contain material that either digresses from the main subject matter of the book or is more sophisticated than what precedes them and may be omitted.

Exercises are found at the ends of chapters. Some (especially those whose results are applied in the text) have solutions at the back of the book. We of course encourage you to try them yourself first!

The Notes at the ends of chapters include references to original papers, suggestions for further reading, and occasionally "complements." These generally contain related material not required elsewhere in the book—sharper versions of lemmas or results that require somewhat greater prerequisites.

The Notation Index at the end of the book lists many recurring symbols.

Much of the book is organized by method, rather than by example. The reader may notice that, in the course of illustrating techniques, we return again and again to certain families of chains—random walks on tori and hypercubes, simple card shuffles, proper colorings of graphs. In our defense we offer an anecdote.

In 1991 one of us (Y. Peres) arrived as a postdoc at Yale and visited Shizuo Kakutani, whose rather large office was full of books and papers, with bookcases and boxes from floor to ceiling. A narrow path led from the door to Kakutani's desk, which was also overflowing with papers. Kakutani admitted that he sometimes had difficulty locating particular papers, but he proudly explained that he had found a way to solve the problem. He would make four or five copies of any really interesting paper and put them in different corners of the office. When searching, he would be sure to find at least one of the copies....

Cross-references in the text and the Index should help you track earlier occurrences of an example. You may also find the chapter dependency diagrams below useful.

We have included brief accounts of some background material in Appendix A. These are intended primarily to set terminology and notation, and we hope you will consult suitable textbooks for unfamiliar material.

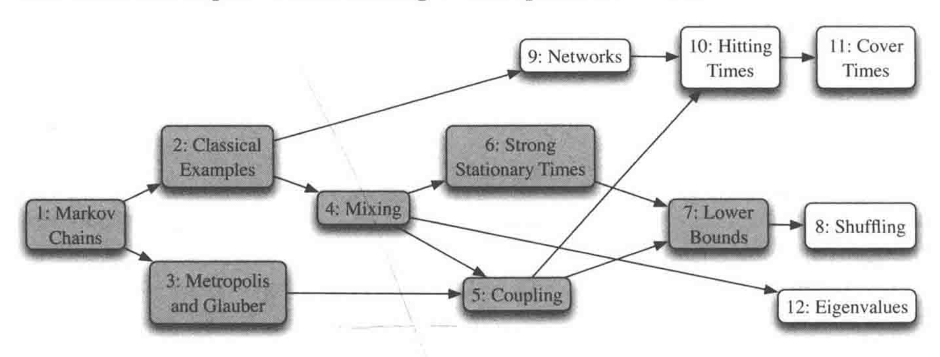
Be aware that we occasionally write symbols representing a real number when an integer is required (see, e.g., the $\binom{n}{\delta k}$'s in the proof of Proposition 13.31). We hope the reader will realize that this omission of floor or ceiling brackets (and the details of analyzing the resulting perturbations) is in her or his best interest as much as it is in ours.

For the Instructor

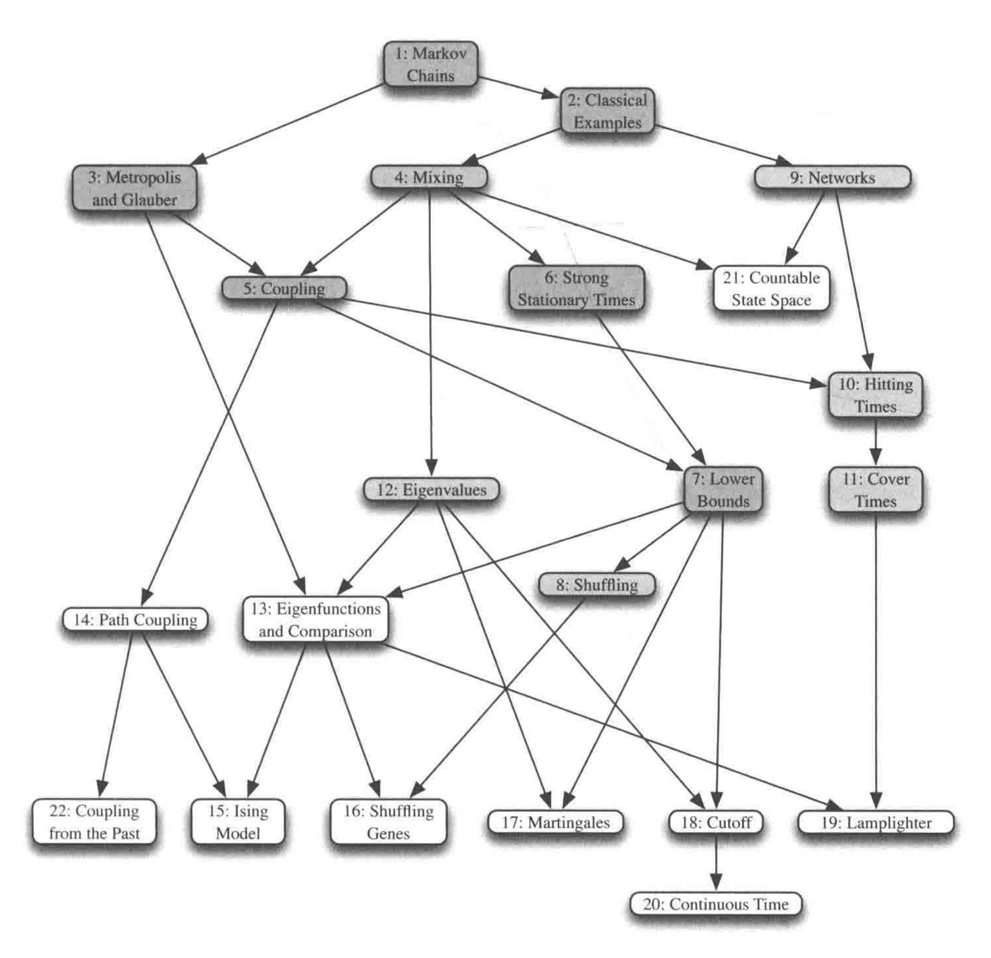
The prerequisites this book demands are a first course in probability, linear algebra, and, inevitably, a certain degree of mathematical maturity. When introducing material which is standard in other undergraduate courses—e.g., groups—we provide definitions, but often hope the reader has some prior experience with the concepts.

In Part I, we have worked hard to keep the material accessible and engaging for students. (Starred sections are more sophisticated and are not required for what follows immediately; they can be omitted.)

Here are the dependencies among the chapters of Part I:



Chapters 1 through 7, shown in gray, form the core material, but there are several ways to proceed afterwards. Chapter 8 on shuffling gives an early rich application but is not required for the rest of Part I. A course with a probabilistic focus might cover Chapters 9, 10, and 11. To emphasize spectral methods and combinatorics, cover Chapters 8 and 12 and perhaps continue on to Chapters 13 and 17.



The logical dependencies of chapters. The core Chapters 1 through 7 are in dark gray, the rest of Part I is in light gray, and Part II is in white.

While our primary focus is on chains with finite state spaces run in discrete time, continuous-time and countable-state-space chains are both discussed—in Chapters 20 and 21, respectively.

We have also included Appendix B, an introduction to simulation methods, to help motivate the study of Markov chains for students with more applied interests. A course leaning towards theoretical computer science and/or statistical mechanics might start with Appendix B, cover the core material, and then move on to Chapters 14, 15, and 22.

Of course, depending on the interests of the instructor and the ambitions and abilities of the students, any of the material can be taught! Above we include a full diagram of dependencies of chapters. Its tangled nature results from the interconnectedness of the area: a given technique can be applied in many situations, while a particular problem may require several techniques for full analysis.

For the Expert

Several other recent books treat Markov chain mixing. Our account is more comprehensive than those of Häggström (2002), Jerrum (2003), or Montenegro and Tetali (2006), yet not as exhaustive as Aldous and Fill (1999). Norris (1998) gives an introduction to Markov chains and their applications, but does not focus on mixing. Since this is a textbook, we have aimed for accessibility and comprehensibility, particularly in Part I.

What is different or novel in our approach to this material?

- Our approach is probabilistic whenever possible. We introduce the random mapping representation of chains early and use it in formalizing randomized stopping times and in discussing grand coupling and evolving sets. We also integrate "classical" material on networks, hitting times, and cover times and demonstrate its usefulness for bounding mixing times.
- We provide an introduction to several major statistical mechanics models, most notably the Ising model, and collect results on them in one place.
- We give expository accounts of several modern techniques and examples, including evolving sets, the cutoff phenomenon, lamplighter chains, and the L-reversal chain.
- We systematically treat lower bounding techniques, including several applications of Wilson's method.
- We use the transportation metric to unify our account of path coupling and draw connections with earlier history.
- We present an exposition of coupling from the past by Propp and Wilson, the originators of the method.

Acknowledgements

The authors thank the Mathematical Sciences Research Institute, the National Science Foundation VIGRE grant to the Department of Statistics at the University of California, Berkeley, and National Science Foundation grants DMS-0244479 and DMS-0104073 for support. We also thank Hugo Rossi for suggesting we embark on this project. Thanks to Blair Ahlquist, Tonci Antunovic, Elisa Celis, Paul Cuff, Jian Ding, Ori Gurel-Gurevich, Tom Hayes, Itamar Landau, Yun Long, Karola Mészáros, Shobhana Murali, Weiyang Ning, Tomoyuki Shirai, Walter Sun, Sithparran Vanniasegaram, and Ariel Yadin for corrections to an earlier version and making valuable suggestions. Yelena Shvets made the illustration in Section 6.5.4. The simulations of the Ising model in Chapter 15 are due to Raissa D'Souza. We thank László Lovász for useful discussions. We are indebted to Alistair Sinclair for his work co-organizing the M.S.R.I. program *Probability, Algorithms, and Statistical Physics* in 2005, where work on this book began. We thank Robert Calhoun for technical assistance.

Finally, we are greatly indebted to David Aldous and Persi Diaconis, who initiated the modern point of view on finite Markov chains and taught us much of what we know about the subject.

Contents

Part I: Basic Methods and Examples

Everything should be made as simple as possible, but not simpler.
–Paraphrase of a quotation from Einstein (1934).

CHAPTER 1

Introduction to Finite Markov Chains

1.1. Finite Markov Chains

A finite Markov chain is a process which moves among the elements of a finite set Ω in the following manner: when at $x \in \Omega$, the next position is chosen according to a fixed probability distribution $P(x, \cdot)$. More precisely, a sequence of random variables $(X_0, X_1, \ldots)$ is a ***Markov chain with state space*** Ω ***and transition matrix*** P if for all $x, y \in \Omega$, all $t \geq 1$, and all events $H_{t-1} = \bigcap_{s=0}^{t-1} \{X_s = x_s\}$ satisfying $\mathbf{P}(H_{t-1} \cap \{X_t = x\}) > 0$, we have

$$\mathbf{P}\left\{X_{t+1} = y \mid H_{t-1} \cap \{X_t = x\}\right\} = \mathbf{P}\left\{X_{t+1} = y \mid X_t = x\right\} = P(x, y). \tag{1.1}$$

Equation (1.1), often called the ***Markov property***, means that the conditional probability of proceeding from state x to state y is the same, no matter what sequence $x_0, x_1, \ldots, x_{t-1}$ of states precedes the current state x. This is exactly why the $|\Omega| \times |\Omega|$ matrix P suffices to describe the transitions.

The x-th row of P is the distribution $P(x, \cdot)$. Thus P is ***stochastic***, that is, its entries are all non-negative and

$$\sum_{y \in \Omega} P(x, y) = 1 \qquad \text{for all } x \in \Omega.$$

EXAMPLE 1.1. A certain frog lives in a pond with two lily pads, *east* and *west*. A long time ago, he found two coins at the bottom of the pond and brought one up to each lily pad. Every morning, the frog decides whether to jump by tossing the current lily pad's coin. If the coin lands heads up, the frog jumps to the other lily pad. If the coin lands tails up, he remains where he is.

Let $\Omega = \{e, w\}$, and let $(X_0, X_1, \ldots)$ be the sequence of lily pads occupied by the frog on Sunday, Monday, ... Given the source of the coins, we should not assume that they are fair! Say the coin on the east pad has probability p of landing



FIGURE 1.1. A randomly jumping frog. Whenever he tosses heads, he jumps to the other lily pad.

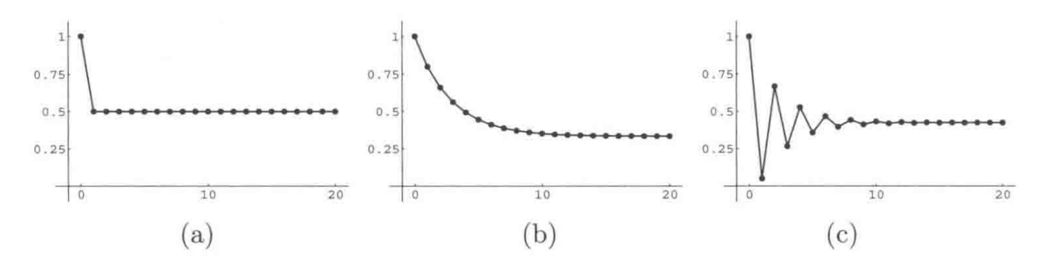


FIGURE 1.2. The probability of being on the east pad (started from the east pad) plotted versus time for (a) $p = q = 1/2$, (b) $p = 0.2$ and $q = 0.1$, (c) $p = 0.95$ and $q = 0.7$. The long-term limiting probabilities are $1/2$, $1/3$, and $14/33 \approx 0.42$, respectively.

heads up, while the coin on the west pad has probability q of landing heads up. The frog's rules for jumping imply that if we set

$$P = \begin{pmatrix} P(e,e) & P(e,w) \\ P(w,e) & P(w,w) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \tag{1.2}$$

then $(X_0, X_1, \dots)$ is a Markov chain with transition matrix P. Note that the first row of P is the conditional distribution of X_{t+1} given that $X_t = e$, while the second row is the conditional distribution of X_{t+1} given that $X_t = w$.

Assume that the frog spends Sunday on the east pad. When he awakens Monday, he has probability p of moving to the west pad and probability $1-p$ of staying on the east pad. That is,

$$\mathbf{P}\{X_1 = e \mid X_0 = e\} = 1 - p, \qquad \mathbf{P}\{X_1 = w \mid X_0 = e\} = p. \tag{1.3}$$

What happens Tuesday? By considering the two possibilities for X_1, we see that

$$\mathbf{P}\{X_2 = e \mid X_0 = e\} = (1-p)(1-p) + pq \tag{1.4}$$

and

$$\mathbf{P}\{X_2 = w \mid X_0 = e\} = (1-p)p + p(1-q). \tag{1.5}$$

While we could keep writing out formulas like (1.4) and (1.5), there is a more systematic approach. We can store our distribution information in a row vector

$$\mu_t := (\mathbf{P}\{X_t = e \mid X_0 = e\}, \mathbf{P}\{X_t = w \mid X_0 = e\}).$$

Our assumption that the frog starts on the east pad can now be written as $\mu_0 = (1, 0)$, while (1.3) becomes $\mu_1 = \mu_0 P$.

Multiplying by P on the right updates the distribution by another step:

$$\mu_t = \mu_{t-1} P \quad \text{for all } t \geq 1. \tag{1.6}$$

Indeed, for any initial distribution μ_0,

$$\mu_t = \mu_0 P^t \quad \text{for all } t \geq 0. \tag{1.7}$$

How does the distribution μ_t behave in the long term? Figure 1.2 suggests that μ_t has a limit π (whose value depends on p and q) as $t \to \infty$. Any such limit distribution π must satisfy

$$\pi = \pi P,$$

which implies (after a little algebra) that

$$\pi(e) = \frac{q}{p+q}, \qquad \pi(w) = \frac{p}{p+q}.$$

If we define

$$\Delta_t = \mu_t(e) - \frac{q}{p+q} \quad \text{for all } t \geq 0,$$

then by the definition of μ_{t+1} the sequence (Δ_t) satisfies

$$\Delta_{t+1} = \mu_t(e)(1-p) + (1-\mu_t(e))(q) - \frac{q}{p+q} = (1-p-q)\Delta_t. \tag{1.8}$$

We conclude that when $0 < p < 1$ and $0 < q < 1$,

$$\lim_{t\to\infty} \mu_t(e) = \frac{q}{p+q} \quad \text{and} \quad \lim_{t\to\infty} \mu_t(w) = \frac{p}{p+q} \tag{1.9}$$

for any initial distribution μ_0. As we suspected, μ_t approaches π as $t \to \infty$.

REMARK 1.2. The traditional theory of finite Markov chains is concerned with convergence statements of the type seen in (1.9), that is, with the rate of convergence as $t \to \infty$ for a *fixed chain*. Note that $1 - p - q$ is an eigenvalue of the frog's transition matrix P. Note also that this eigenvalue determines the rate of convergence in (1.9), since by (1.8) we have

$$\Delta_t = (1-p-q)^t \Delta_0.$$

The computations we just did for a two-state chain generalize to any finite Markov chain. In particular, the distribution at time t can be found by matrix multiplication. Let $(X_0, X_1, \dots)$ be a finite Markov chain with state space Ω and transition matrix P, and let the row vector μ_t be the distribution of X_t:

$$\mu_t(x) = \mathbf{P}\{X_t = x\} \quad \text{for all } x \in \Omega.$$

By conditioning on the possible predecessors of the $(t+1)$-st state, we see that

$$\mu_{t+1}(y) = \sum_{x\in\Omega} \mathbf{P}\{X_t = x\}P(x,y) = \sum_{x\in\Omega} \mu_t(x)P(x,y) \quad \text{for all } y \in \Omega.$$

Rewriting this in vector form gives

$$\mu_{t+1} = \mu_t P \quad \text{for } t \geq 0$$

and hence

$$\mu_t = \mu_0 P^t \quad \text{for } t \geq 0. \tag{1.10}$$

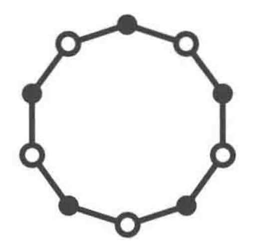
Since we will often consider Markov chains with the same transition matrix but different starting distributions, we introduce the notation $\mathbf{P}_\mu$ and $\mathbf{E}_\mu$ for probabilities and expectations given that $\mu_0 = \mu$. Most often, the initial distribution will be concentrated at a single definite starting state x. We denote this distribution by δ_x:

$$\delta_x(y) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{if } y \neq x. \end{cases}$$

We write simply $\mathbf{P}_x$ and $\mathbf{E}_x$ for $\mathbf{P}_{\delta_x}$ and $\mathbf{E}_{\delta_x}$, respectively.

These definitions and (1.10) together imply that

$$\mathbf{P}_x\{X_t = y\} = (\delta_x P^t)(y) = P^t(x,y).$$



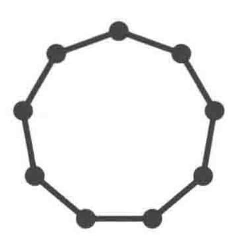


FIGURE 1.3. Random walk on $\mathbb{Z}_{10}$ is periodic, since every step goes from an even state to an odd state, or vice-versa. Random walk on $\mathbb{Z}_9$ is aperiodic.

That is, the probability of moving in t steps from x to y is given by the (x, y)-th entry of P^t. We call these entries the ***t-step transition probabilities***.

NOTATION. A probability distribution μ on Ω will be identified with a row vector. For any event $A \subset \Omega$, we write

$$\pi(A) = \sum_{x \in A} \mu(x).$$

For $x \in \Omega$, the row of P indexed by x will be denoted by $P(x, \cdot)$.

REMARK 1.3. The way we constructed the matrix P has forced us to treat distributions as row vectors. In general, if the chain has distribution μ at time t, then it has distribution μP at time $t + 1$. *Multiplying a row vector by P on the right takes you from today's distribution to tomorrow's distribution.*

What if we multiply a column vector f by P on the left? Think of f as a function on the state space Ω (for the frog of Example 1.1, we might take $f(x)$ to be the area of the lily pad x). Consider the x-th entry of the resulting vector:

$$Pf(x) = \sum_y P(x, y) f(y) = \sum_y f(y) \mathbf{P}_x\{X_1 = y\} = \mathbf{E}_x(f(X_1)).$$

That is, the x-th entry of Pf tells us the expected value of the function f at tomorrow's state, given that we are at state x today. *Multiplying a column vector by P on the left takes us from a function on the state space to the expected value of that function tomorrow.*

1.2. Random Mapping Representation

We begin this section with an example.

EXAMPLE 1.4 (Random walk on the n-cycle). Let $\Omega = \mathbb{Z}_n = \{0, 1, \ldots, n-1\}$, the set of remainders modulo n. Consider the transition matrix

$$P(j, k) = \begin{cases} 1/2 & \text{if } k \equiv j + 1 \pmod{n}, \\ 1/2 & \text{if } k \equiv j - 1 \pmod{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.11}$$

The associated Markov chain (X_t) is called ***random walk on the n-cycle***. The states can be envisioned as equally spaced dots arranged in a circle (see Figure 1.3).

Rather than writing down the transition matrix in (1.11), this chain can be specified simply in words: at each step, a coin is tossed. If the coin lands heads up, the walk moves one step clockwise. If the coin lands tails up, the walk moves one step counterclockwise.

More precisely, suppose that Z is a random variable which is equally likely to take on the values -1 and $+1$. If the current state of the chain is $j \in \mathbb{Z}_n$, then the next state is $j + Z \bmod n$. For any $k \in \mathbb{Z}_n$,

$$\mathbf{P}\{(j + Z) \bmod n = k\} = P(j, k).$$

In other words, the distribution of $(j + Z) \bmod n$ equals $P(j, \cdot)$.

A ***random mapping representation*** of a transition matrix P on state space Ω is a function $f : \Omega \times \Lambda \to \Omega$, along with a Λ-valued random variable Z, satisfying

$$\mathbf{P}\{f(x, Z) = y\} = P(x, y).$$

The reader should check that if $Z_1, Z_2, \ldots$ is a sequence of independent random variables, each having the same distribution as Z, and X_0 has distribution μ, then the sequence $(X_0, X_1, \ldots)$ defined by

$$X_n = f(X_{n-1}, Z_n) \quad \text{for } n \geq 1$$

is a Markov chain with transition matrix P and initial distribution μ.

For the example of the simple random walk on the cycle, setting $\Lambda = \{1, -1\}$, each Z_i uniform on Λ, and $f(x, z) = x + z \bmod n$ yields a random mapping representation.

PROPOSITION 1.5. *Every transition matrix on a finite state space has a random mapping representation.*

PROOF. Let P be the transition matrix of a Markov chain with state space $\Omega = \{x_1, \ldots, x_n\}$. Take $\Lambda = [0, 1]$; our auxiliary random variables $Z, Z_1, Z_2, \ldots$ will be uniformly chosen in this interval. Set $F_{j,k} = \sum_{i=1}^{k} P(x_j, x_i)$ and define

$$f(x_j, z) := x_k \text{ when } F_{j,k-1} < z \leq F_{j,k}.$$

We have

$$\mathbf{P}\{f(x_j, Z) = x_k\} = \mathbf{P}\{F_{j,k-1} < Z \leq F_{j,k}\} = P(x_j, x_k).$$

■

Note that, unlike transition matrices, random mapping representations are far from unique. For instance, replacing the function $f(x, z)$ in the proof of Proposition 1.5 with $f(x, 1 - z)$ yields a different representation of the same transition matrix.

Random mapping representations are crucial for simulating large chains. They can also be the most convenient way to describe a chain. We will often give rules for how a chain proceeds from state to state, using some extra randomness to determine where to go next; such discussions are implicit random mapping representations. Finally, random mapping representations provide a way to coordinate two (or more) chain trajectories, as we can simply use the same sequence of auxiliary random variables to determine updates. This technique will be exploited in Chapter 5, on coupling Markov chain trajectories, and elsewhere.

1.3. Irreducibility and Aperiodicity

We now make note of two simple properties possessed by most interesting chains. Both will turn out to be necessary for the Convergence Theorem (Theorem 4.9) to be true.

A chain P is called ***irreducible*** if for any two states $x, y \in \Omega$ there exists an integer t (possibly depending on x and y) such that $P^t(x, y) > 0$. This means that it is possible to get from any state to any other state using only transitions of positive probability. We will generally assume that the chains under discussion are irreducible. (Checking that specific chains are irreducible can be quite interesting; see, for instance, Section 2.6 and Example B.5. See Section 1.7 for a discussion of all the ways in which a Markov chain can fail to be irreducible.)

Let $\mathcal{T}(x) := \{t \geq 1 : P^t(x, x) > 0\}$ be the set of times when it is possible for the chain to return to starting position x. The ***period*** of state x is defined to be the greatest common divisor of $\mathcal{T}(x)$.

LEMMA 1.6. *If P is irreducible, then* $\gcd \mathcal{T}(x) = \gcd \mathcal{T}(y)$ *for all $x, y \in \Omega$.*

PROOF. Fix two states x and y. There exist non-negative integers r and ℓ such that $P^r(x, y) > 0$ and $P^\ell(y, x) > 0$. Letting $m = r + \ell$, we have $m \in \mathcal{T}(x) \cap \mathcal{T}(y)$ and $\mathcal{T}(x) \subset \mathcal{T}(y) - m$, whence $\gcd \mathcal{T}(y)$ divides all elements of $\mathcal{T}(x)$. We conclude that $\gcd \mathcal{T}(y) \leq \gcd \mathcal{T}(x)$. By an entirely parallel argument, $\gcd \mathcal{T}(x) \leq \gcd \mathcal{T}(y)$. ∎

For an irreducible chain, the period of the chain is defined to be the period which is common to all states. The chain will be called ***aperiodic*** if all states have period 1. If a chain is not aperiodic, we call it ***periodic***.

PROPOSITION 1.7. *If P is aperiodic and irreducible, then there is an integer r such that $P^r(x, y) > 0$ for all $x, y \in \Omega$.*

PROOF. We use the following number-theoretic fact: any set of non-negative integers which is closed under addition and which has greatest common divisor 1 must contain all but finitely many of the non-negative integers. (See Lemma 1.27 in the Notes of this chapter for a proof.) For $x \in \Omega$, recall that $\mathcal{T}(x) = \{t \geq 1 : P^t(x, x) > 0\}$. Since the chain is aperiodic, the gcd of $\mathcal{T}(x)$ is 1. The set $\mathcal{T}(x)$ is closed under addition: if $s, t \in \mathcal{T}(x)$, then $P^{s+t}(x, x) \geq P^s(x, x)P^t(x, x) > 0$, and hence $s + t \in \mathcal{T}(x)$. Therefore there exists a $t(x)$ such that $t \geq t(x)$ implies $t \in \mathcal{T}(x)$. By irreducibility we know that for any $y \in \Omega$ there exists $r = r(x, y)$ such that $P^r(x, y) > 0$. Therefore, for $t \geq t(x) + r$,

$$P^t(x, y) \geq P^{t-r}(x, x)P^r(x, y) > 0.$$

For $t \geq t'(x) := t(x) + \max_{y \in \Omega} r(x, y)$, we have $P^t(x, y) > 0$ for all $y \in \Omega$. Finally, if $t \geq \max_{x \in \Omega} t'(x)$, then $P^t(x, y) > 0$ for all $x, y \in \Omega$. ∎

Suppose that a chain is irreducible with period two, e.g. the simple random walk on a cycle of even length (see Figure 1.3). The state space Ω can be partitioned into two classes, say ***even*** and ***odd***, such that the chain makes transitions only between states in complementary classes. (Exercise 1.6 examines chains with period b.)

Let P have period two, and suppose that x_0 is an even state. The probability distribution of the chain after $2t$ steps, $P^{2t}(x_0, \cdot)$, is supported on even states, while the distribution of the chain after $2t + 1$ steps is supported on odd states. It is evident that we cannot expect the distribution $P^t(x_0, \cdot)$ to converge as $t \to \infty$.

Fortunately, a simple modification can repair periodicity problems. Given an arbitrary transition matrix P, let $Q = \frac{I+P}{2}$ (here I is the $|\Omega| \times |\Omega|$ identity matrix). (One can imagine simulating Q as follows: at each time step, flip a fair coin. If it comes up heads, take a step in P; if tails, then stay at the current state.) Since $Q(x,x) > 0$ for all $x \in \Omega$, the transition matrix Q is aperiodic. We call Q a ***lazy version of*** P. It will often be convenient to analyze lazy versions of chains.

EXAMPLE 1.8 (The n-cycle, revisited). Recall random walk on the n-cycle, defined in Example 1.4. For every $n \geq 1$, random walk on the n-cycle is irreducible.

Random walk on any even-length cycle is periodic, since $\gcd\{t : P^t(x,x) > 0\} = 2$ (see Figure 1.3). Random walk on an odd-length cycle is aperiodic.

The transition matrix Q for lazy random walk on the n-cycle is

$$Q(j,k) = \begin{cases} 1/4 & \text{if } k \equiv j+1 \pmod n, \\ 1/2 & \text{if } k \equiv j \pmod n, \\ 1/4 & \text{if } k \equiv j-1 \pmod n, \\ 0 & \text{otherwise.} \end{cases} \tag{1.12}$$

Lazy random walk on the n-cycle is both irreducible and aperiodic for every n.

REMARK 1.9. Establishing that a Markov chain is irreducible is not always trivial; see Example B.5, and also Thurston (1990).

1.4. Random Walks on Graphs

Random walk on the n-cycle, which is shown in Figure 1.3, is a simple case of an important type of Markov chain.

A ***graph*** $G = (V, E)$ consists of a ***vertex set*** V and an ***edge set*** E, where the elements of E are unordered pairs of vertices: $E \subset \{\{x,y\} : x, y \in V, x \neq y\}$. We can think of V as a set of dots, where two dots x and y are joined by a line if and only if $\{x,y\}$ is an element of the edge set. When $\{x,y\} \in E$, we write $x \sim y$ and say that y is a ***neighbor*** of x (and also that x is a neighbor of y). The ***degree*** $\deg(x)$ of a vertex x is the number of neighbors of x.

Given a graph $G = (V, E)$, we can define ***simple random walk on*** G to be the Markov chain with state space V and transition matrix

$$P(x,y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y \sim x, \\ 0 & \text{otherwise.} \end{cases} \tag{1.13}$$

That is to say, when the chain is at vertex x, it examines all the neighbors of x, picks one uniformly at random, and moves to the chosen vertex.

EXAMPLE 1.10. Consider the graph G shown in Figure 1.4. The transition matrix of simple random walk on G is

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

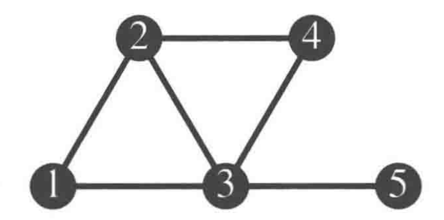


FIGURE 1.4. An example of a graph with vertex set $\{1, 2, 3, 4, 5\}$ and 6 edges.

REMARK 1.11. We have chosen a narrow definition of "graph" for simplicity. It is sometimes useful to allow edges connecting a vertex to itself, called ***loops***. It is also sometimes useful to allow multiple edges connecting a single pair of vertices. Loops and multiple edges both contribute to the degree of a vertex and are counted as options when a simple random walk chooses a direction. See Section 6.5.1 for an example.

We will have much more to say about random walks on graphs throughout this book—but especially in Chapter 9.

1.5. Stationary Distributions

1.5.1. Definition. We saw in Example 1.1 that a distribution π on Ω satisfying

$$\pi = \pi P \tag{1.14}$$

can have another interesting property: in that case, π was the long-term limiting distribution of the chain. We call a probability π satisfying (1.14) a ***stationary distribution*** of the Markov chain. Clearly, if π is a stationary distribution and $\mu_0 = \pi$ (i.e. the chain is started in a stationary distribution), then $\mu_t = \pi$ for all $t \geq 0$.

Note that we can also write (1.14) elementwise. An equivalent formulation is

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(x, y) \quad \text{for all } y \in \Omega. \tag{1.15}$$

EXAMPLE 1.12. Consider simple random walk on a graph $G = (V, E)$. For any vertex $y \in V$,

$$\sum_{x \in V} \deg(x) P(x, y) = \sum_{x \sim y} \frac{\deg(x)}{\deg(x)} = \deg(y). \tag{1.16}$$

To get a probability, we simply normalize by $\sum_{y \in V} \deg(y) = 2|E|$ (a fact the reader should check). We conclude that the probability measure

$$\pi(y) = \frac{\deg(y)}{2|E|} \quad \text{for all } y \in \Omega,$$

which is proportional to the degrees, is always a stationary distribution for the walk. For the graph in Figure 1.4,

$$\pi = \left(\tfrac{2}{12}, \tfrac{3}{12}, \tfrac{4}{12}, \tfrac{2}{12}, \tfrac{1}{12}\right).$$

If G has the property that every vertex has the same degree d, we call G ***d-regular***. In this case $2|E| = d|V|$ and the uniform distribution $\pi(y) = 1/|V|$ for every $y \in V$ is stationary.

A central goal of this chapter and of Chapter 4 is to prove a general yet precise version of the statement that "finite Markov chains converge to their stationary distributions." Before we can analyze the time required to be close to stationarity, we must be sure that it is finite! In this section we show that, under mild restrictions, stationary distributions exist and are unique. Our strategy of building a candidate distribution, then verifying that it has the necessary properties, may seem cumbersome. However, the tools we construct here will be applied in many other places. In Section 4.3, we will show that irreducible and aperiodic chains do, in fact, converge to their stationary distributions in a precise sense.

1.5.2. Hitting and first return times. Throughout this section, we assume that the Markov chain $(X_0, X_1, \dots)$ under discussion has finite state space Ω and transition matrix P. For $x \in \Omega$, define the ***hitting time*** for x to be

$$\tau_x := \min\{t \geq 0 : X_t = x\},$$

the first time at which the chain visits state x. For situations where only a visit to x at a positive time will do, we also define

$$\tau_x^+ := \min\{t \geq 1 \, : \, X_t = x\}.$$

When $X_0 = x$, we call τ_x^+ the ***first return time***.

LEMMA 1.13. *For any states x and y of an irreducible chain, $\mathbf{E}_x(\tau_y^+) < \infty$.*

PROOF. The definition of irreducibility implies that there exist an integer $r > 0$ and a real $\varepsilon > 0$ with the following property: for any states $z, w \in \Omega$, there exists a $j \leq r$ with $P^j(z, w) > \varepsilon$. Thus for any value of X_t, the probability of hitting state y at a time between t and $t + r$ is at least ε. Hence for $k > 0$ we have

$$\mathbf{P}_x\{\tau_y^+ > kr\} \leq (1-\varepsilon)\mathbf{P}_x\{\tau_y^+ > (k-1)r\}. \tag{1.17}$$

Repeated application of (1.17) yields

$$\mathbf{P}_x\{\tau_y^+ > kr\} \leq (1-\varepsilon)^k. \tag{1.18}$$

Recall that when Y is a non-negative integer-valued random variable, we have

$$\mathbf{E}(Y) = \sum_{t \geq 0} \mathbf{P}\{Y > t\}.$$

Since $\mathbf{P}_x\{\tau_y^+ > t\}$ is a decreasing function of t, (1.18) suffices to bound all terms of the corresponding expression for $\mathbf{E}_x(\tau_y^+)$:

$$\mathbf{E}_x(\tau_y^+) = \sum_{t \geq 0} \mathbf{P}_x\{\tau_y^+ > t\} \leq \sum_{k \geq 0} r\mathbf{P}_x\{\tau_y^+ > kr\} \leq r \sum_{k \geq 0} (1-\varepsilon)^k < \infty.$$

■

1.5.3. Existence of a stationary distribution. The Convergence Theorem (Theorem 4.9 below) implies that the "long-term" fractions of time a finite irreducible aperiodic Markov chain spends in each state coincide with the chain's stationary distribution. However, we have not yet demonstrated that stationary distributions exist! To build a candidate distribution, we consider a sojourn of the chain from some arbitrary state z back to z. Since visits to z break up the trajectory of the chain into identically distributed segments, it should not be surprising that the average fraction of time per segment spent in each state y coincides with the "long-term" fraction of time spent in y.

PROPOSITION 1.14. *Let P be the transition matrix of an irreducible Markov chain. Then*

(i) *there exists a probability distribution π on Ω such that $\pi = \pi P$ and $\pi(x) > 0$ for all $x \in \Omega$, and moreover,*
(ii) $\pi(x) = \frac{1}{\mathbf{E}_x(\tau_x^+)}$.

REMARK 1.15. We will see in Section 1.7 that existence of π does not need irreducibility, but positivity does.

PROOF. Let $z \in \Omega$ be an arbitrary state of the Markov chain. We will closely examine the time the chain spends, on average, at each state in between visits to z. Hence define

$$\begin{aligned}\tilde{\pi}(y) &:= \mathbf{E}_z(\text{number of visits to } y \text{ before returning to } z) \\ &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ > t\}.\end{aligned} \tag{1.19}$$

For any state y, we have $\tilde{\pi}(y) \le \mathbf{E}_z \tau_z^+$. Hence Lemma 1.13 ensures that $\tilde{\pi}(y) < \infty$ for all $y \in \Omega$. We check that $\tilde{\pi}$ is stationary, starting from the definition:

$$\sum_{x\in\Omega} \tilde{\pi}(x) P(x,y) = \sum_{x\in\Omega}\sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = x, \tau_z^+ > t\} P(x,y). \tag{1.20}$$

Because the event $\{\tau_z^+ \ge t+1\} = \{\tau_z^+ > t\}$ is determined by $X_0, \ldots, X_t$,

$$\mathbf{P}_z\{X_t = x,\, X_{t+1} = y,\, \tau_z^+ \ge t+1\} = \mathbf{P}_z\{X_t = x,\, \tau_z^+ \ge t+1\} P(x,y). \tag{1.21}$$

Reversing the order of summation in (1.20) and using the identity (1.21) shows that

$$\begin{aligned}\sum_{x\in\Omega} \tilde{\pi}(x) P(x,y) &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_{t+1} = y, \tau_z^+ \ge t+1\} \\ &= \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ \ge t\}.\end{aligned} \tag{1.22}$$

The expression in (1.22) is very similar to (1.19), so we are almost done. In fact,

$$\begin{aligned}
&\sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ \geq t\} \\
&\quad = \tilde{\pi}(y) - \mathbf{P}_z\{X_0 = y, \tau_z^+ > 0\} + \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ = t\} \\
&\quad = \tilde{\pi}(y) - \mathbf{P}_z\{X_0 = y\} + \mathbf{P}_z\{X_{\tau_z^+} = y\}. &(1.23)\\
&\quad = \tilde{\pi}(y). &(1.24)
\end{aligned}$$

The equality (1.24) follows by considering two cases:

$y = z$: Since $X_0 = z$ and $X_{\tau_z^+} = z$, the last two terms of (1.23) are both 1, and they cancel each other out.

$y \neq z$: Here both terms of (1.23) are 0.

Therefore, combining (1.22) with (1.24) shows that $\tilde{\pi} = \tilde{\pi}P$.

Finally, to get a probability measure, we normalize by $\sum_x \tilde{\pi}(x) = \mathbf{E}_z(\tau_z^+)$:

$$\pi(x) = \frac{\tilde{\pi}(x)}{\mathbf{E}_z(\tau_z^+)} \quad \text{satisfies } \pi = \pi P. \tag{1.25}$$

In particular, for any $x \in \Omega$,

$$\pi(x) = \frac{1}{\mathbf{E}_x(\tau_x^+)}. \tag{1.26}$$

■

The computation at the heart of the proof of Proposition 1.14 can be generalized. A ***stopping time*** τ for (X_t) is a $\{0, 1, \ldots,\} \cup \{\infty\}$-valued random variable such that, for each t, the event $\{\tau = t\}$ is determined by $X_0, \ldots, X_t$. (Stopping times are discussed in detail in Section 6.2.1.) If a stopping time τ replaces τ_z^+ in the definition (1.19) of $\tilde{\pi}$, then the proof that $\tilde{\pi}$ satisfies $\tilde{\pi} = \tilde{\pi}P$ works, provided that τ satisfies both $\mathbf{P}_z\{\tau < \infty\} = 1$ and $\mathbf{P}_z\{X_\tau = z\} = 1$.

If τ is a stopping time, then an immediate consequence of the definition and the Markov property is

$$\begin{aligned}
\mathbf{P}_{x_0}\{(X_{\tau+1}, X_{\tau+2}, \ldots, X_\ell) \in A \mid \tau = k \text{ and } (X_1, \ldots, X_k) = (x_1, \ldots, x_k)\} \\
= \mathbf{P}_{x_k}\{(X_1, \ldots, X_\ell) \in A\}, \quad (1.27)
\end{aligned}$$

for any $A \subset \Omega^\ell$. This is referred to as the ***strong Markov property***. Informally, we say that the chain "starts afresh" at a stopping time. While this is an easy fact for countable state space, discrete-time Markov chains, establishing it for processes in the continuum is more subtle.

1.5.4. Uniqueness of the stationary distribution. Earlier this chapter we pointed out the difference between multiplying a row vector by P on the right and a column vector by P on the left: the former advances a distribution by one step of the chain, while the latter gives the expectation of a function on states, one step of the chain later. We call distributions invariant under right multiplication by P ***stationary***. What about functions that are invariant under left multiplication?

Call a function $h : \Omega \to \mathbb{R}$ ***harmonic at*** x if

$$h(x) = \sum_{y \in \Omega} P(x, y) h(y). \tag{1.28}$$

A function is ***harmonic on*** $D \subset \Omega$ if it is harmonic at every state $x \in D$. If h is regarded as a column vector, then a function which is harmonic on all of Ω satisfies the matrix equation $Ph = h$.

LEMMA 1.16. *Suppose that P is irreducible. A function h which is harmonic at every point of Ω is constant.*

PROOF. Since Ω is finite, there must be a state x_0 such that $h(x_0) = M$ is maximal. If for some state z such that $P(x_0, z) > 0$ we have $h(z) < M$, then

$$h(x_0) = P(x_0, z)h(z) + \sum_{y \neq z} P(x_0, y)h(y) < M, \tag{1.29}$$

a contradiction. It follows that $h(z) = M$ for all states z such that $P(x_0, z) > 0$.

For any $y \in \Omega$, irreducibility implies that there is a sequence $x_0, x_1, \ldots, x_n = y$ with $P(x_i, x_{i+1}) > 0$. Repeating the argument above tells us that $h(y) = h(x_{n-1}) = \cdots = h(x_0) = M$. Thus h is constant. ■

COROLLARY 1.17. *Let P be the transition matrix of an irreducible Markov chain. There exists a unique probability distribution π satisfying $\pi = \pi P$.*

PROOF. By Proposition 1.14 there exists at least one such measure. Lemma 1.16 implies that the kernel of $P - I$ has dimension 1, so the column rank of $P - I$ is $|\Omega| - 1$. Since the row rank of any square matrix is equal to its column rank, the row-vector equation $\nu = \nu P$ also has a one-dimensional space of solutions. This space contains only one vector whose entries sum to 1. ■

REMARK 1.18. Another proof of Corollary 1.17 follows from the Convergence Theorem (Theorem 4.9, proved below). Another simple direct proof is suggested in Exercise 1.13.

1.6. Reversibility and Time Reversals

Suppose a probability π on Ω satisfies

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad \text{for all } x, y \in \Omega. \tag{1.30}$$

The equations (1.30) are called the ***detailed balance equations***.

PROPOSITION 1.19. *Let P be the transition matrix of a Markov chain with state space Ω. Any distribution π satisfying the detailed balance equations* (1.30) *is stationary for P.*

PROOF. Sum both sides of (1.30) over all y:

$$\sum_{y \in \Omega} \pi(y)P(y, x) = \sum_{y \in \Omega} \pi(x)P(x, y) = \pi(x),$$

since P is stochastic. ■

Checking detailed balance is often the simplest way to verify that a particular distribution is stationary. Furthermore, when (1.30) holds,

$$\pi(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n) = \pi(x_n)P(x_n, x_{n-1}) \cdots P(x_1, x_0). \tag{1.31}$$

We can rewrite (1.31) in the following suggestive form:

$$\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_n = x_n\} = \mathbf{P}_\pi\{X_0 = x_n, X_1 = x_{n-1}, \ldots, X_n = x_0\}. \tag{1.32}$$

In other words, if a chain (X_t) satisfies (1.30) and has stationary initial distribution, then the distribution of $(X_0, X_1, \ldots, X_n)$ is the same as the distribution of $(X_n, X_{n-1}, \ldots, X_0)$. For this reason, a chain satisfying (1.30) is called ***reversible***.

EXAMPLE 1.20. Consider the simple random walk on a graph G. We saw in Example 1.12 that the distribution $\pi(x) = \deg(x)/2|E|$ is stationary.

Since

$$\pi(x)P(x,y) = \frac{\deg(x)}{2|E|}\frac{\mathbf{1}_{\{x\sim y\}}}{\deg(x)} = \frac{\mathbf{1}_{\{x\sim y\}}}{2|E|} = \pi(y)P(x,y),$$

the chain is reversible. (Note: here the notation $\mathbf{1}_A$ represents the ***indicator function*** of a set A, for which $\mathbf{1}_A(a) = 1$ if and only if $a \in A$; otherwise $\mathbf{1}_A(a) = 0$.)

EXAMPLE 1.21. Consider the ***biased random walk on the n-cycle***: a particle moves clockwise with probability p and moves counterclockwise with probability $q = 1 - p$.

The stationary distribution remains uniform: if $\pi(k) = 1/n$, then

$$\sum_{j\in\mathbb{Z}_n} \pi(j)P(j,k) = \pi(k-1)p + \pi(k+1)q = \frac{1}{n},$$

whence π is the stationary distribution. However, if $p \neq 1/2$, then

$$\pi(k)P(k,k+1) = \frac{p}{n} \neq \frac{q}{n} = \pi(k+1)P(k+1,k).$$

The ***time reversal*** of an irreducible Markov chain with transition matrix P and stationary distribution π is the chain with matrix

$$\widehat{P}(x,y) := \frac{\pi(y)P(y,x)}{\pi(x)}. \tag{1.33}$$

The stationary equation $\pi = \pi P$ implies that $\widehat{P}$ is a stochastic matrix. Proposition 1.22 shows that the terminology "time reversal" is deserved.

PROPOSITION 1.22. *Let (X_t) be an irreducible Markov chain with transition matrix P and stationary distribution π. Write $(\widehat{X}_t)$ for the time-reversed chain with transition matrix $\widehat{P}$. Then π is stationary for $\widehat{P}$, and for any $x_0, \ldots, x_t \in \Omega$ we have*

$$\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_t = x_t\} = \mathbf{P}_\pi\{\widehat{X}_0 = x_t, \ldots, \widehat{X}_t = x_0\}.$$

PROOF. To check that π is stationary for $\widehat{P}$, we simply compute

$$\sum_{y\in\Omega} \pi(y)\widehat{P}(y,x) = \sum_{y\in\Omega} \pi(y)\frac{\pi(x)P(x,y)}{\pi(y)} = \pi(x).$$

To show the probabilities of the two trajectories are equal, note that

$$\begin{aligned}
\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_n = x_n\} &= \pi(x_0)P(x_0,x_1)P(x_1,x_2)\cdots P(x_{n-1},x_n) \\
&= \pi(x_n)\widehat{P}(x_n,x_{n-1})\cdots\widehat{P}(x_2,x_1)\widehat{P}(x_1,x_0) \\
&= \mathbf{P}_\pi\{\hat{X}_0 = x_n, \ldots, \hat{X}_n = x_0\},
\end{aligned}$$

since $P(x_{i-1},x_i) = \pi(x_i)\widehat{P}(x_i,x_{i-1})/\pi(x_{i-1})$ for each i. ■

Observe that if a chain with transition matrix P is reversible, then $\widehat{P} = P$.

1.7. Classifying the States of a Markov Chain*

We will occasionally need to study chains which are *not* irreducible—see, for instance, Sections 2.1, 2.2 and 2.4. In this section we describe a way to classify the states of a Markov chain. This classification clarifies what can occur when irreducibility fails.

Let P be the transition matrix of a Markov chain on a finite state space Ω. Given $x, y \in \Omega$, we say that y ***is accessible from*** x and write $x \to y$ if there exists an $r > 0$ such that $P^r(x,y) > 0$. That is, $x \to y$ if it is possible for the chain to move from x to y in a finite number of steps. Note that if $x \to y$ and $y \to z$, then $x \to z$.

A state $x \in \Omega$ is called ***essential*** if for all y such that $x \to y$ it is also true that $y \to x$. A state $x \in \Omega$ is ***inessential*** if it is not essential.

We say that x ***communicates with*** y and write $x \leftrightarrow y$ if and only if $x \to y$ and $y \to x$. The equivalence classes under $\leftrightarrow$ are called ***communicating classes***. For $x \in \Omega$, the communicating class of x is denoted by $[x]$.

Observe that when P is irreducible, all the states of the chain lie in a single communicating class.

LEMMA 1.23. *If x is an essential state and $x \to y$, then y is essential.*

PROOF. If $y \to z$, then $x \to z$. Therefore, because x is essential, $z \to x$, whence $z \to y$. ∎

It follows directly from the above lemma that the states in a single communicating class are either all essential or all inessential. We can therefore classify the communicating classes as either essential or inessential.

If $[x] = \{x\}$ and x is inessential, then once the chain leaves x, it never returns. If $[x] = \{x\}$ and x is essential, then the chain never leaves x once it first visits x; such states are called ***absorbing***.

LEMMA 1.24. *Every finite chain has at least one essential class.*

PROOF. Define inductively a sequence $(y_0, y_1, \ldots)$ as follows: Fix an arbitrary initial state y_0. For $k \geq 1$, given $(y_0, \ldots, y_{k-1})$, if y_{k-1} is essential, stop. Otherwise, find y_k such that $y_{k-1} \to y_k$ but $y_k \not\to y_{k-1}$.

There can be no repeated states in this sequence, because if $j < k$ and $y_k \to y_j$, then $y_k \to y_{k-1}$, a contradiction.

Since the state space is finite and the sequence cannot repeat elements, it must eventually terminate in an essential state. ∎

Note that a transition matrix P restricted to an essential class $[x]$ is stochastic. That is, $\sum_{y \in [x]} P(x,y) = 1$, since $P(x,z) = 0$ for $z \notin [x]$.

PROPOSITION 1.25. *If π is stationary for the finite transition matrix P, then $\pi(y_0) = 0$ for all inessential states y_0.*

PROOF. Let $\mathcal{C}$ be an essential communicating class. Then

$$\pi P(\mathcal{C}) = \sum_{z \in \mathcal{C}} (\pi P)(z) = \sum_{z \in \mathcal{C}} \left[\sum_{y \in \mathcal{C}} \pi(y) P(y,z) + \sum_{y \notin \mathcal{C}} \pi(y) P(y,z) \right].$$

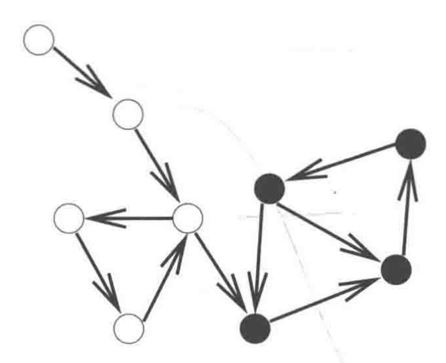


FIGURE 1.5. The directed graph associated to a Markov chain. A directed edge is placed between v and w if and only if $P(v,w) > 0$. Here there is one essential class, which consists of the filled vertices.

We can interchange the order of summation in the first sum, obtaining

$$\pi P(\mathcal{C}) = \sum_{y\in\mathcal{C}} \pi(y) \sum_{z\in\mathcal{C}} P(y,z) + \sum_{z\in\mathcal{C}} \sum_{y\notin\mathcal{C}} \pi(y)P(y,z).$$

For $y \in \mathcal{C}$ we have $\sum_{z\in\mathcal{C}} P(y,z) = 1$, so

$$\pi P(\mathcal{C}) = \pi(\mathcal{C}) + \sum_{z\in\mathcal{C}} \sum_{y\notin\mathcal{C}} \pi(y)P(y,z). \tag{1.34}$$

Since π is invariant, $\pi P(\mathcal{C}) = \pi(\mathcal{C})$. In view of (1.34) we must have $\pi(y)P(y,z) = 0$ for all $y \notin \mathcal{C}$ and $z \in \mathcal{C}$.

Suppose that y_0 is inessential. The proof of Lemma 1.24 shows that there is a sequence of states $y_0, y_1, y_2, \ldots, y_r$ satisfying $P(y_{i-1}, y_i) > 0$, the states $y_0, y_1, \ldots, y_{r-1}$ are inessential, and $y_r \in \mathcal{C}$, where $\mathcal{C}$ is an essential communicating class. Since $P(y_{r-1}, y_r) > 0$ and we just proved that $\pi(y_{r-1})P(y_{r-1}, y_r) = 0$, it follows that $\pi(y_{r-1}) = 0$. If $\pi(y_k) = 0$, then

$$0 = \pi(y_k) = \sum_{y\in\Omega} \pi(y)P(y, y_k).$$

This implies $\pi(y)P(y, y_k) = 0$ for all y. In particular, $\pi(y_{k-1}) = 0$. By induction backwards along the sequence, we find that $\pi(y_0) = 0$. ∎

Finally, we conclude with the following proposition:

PROPOSITION 1.26. *The stationary distribution π for a transition matrix P is unique if and only if there is a unique essential communicating class.*

PROOF. Suppose that there is a unique essential communicating class $\mathcal{C}$. We write $P_{|\mathcal{C}}$ for the restriction of the matrix P to the states in $\mathcal{C}$. Suppose $x \in \mathcal{C}$ and $P(x,y) > 0$. Then since x is essential and $x \to y$, it must be that $y \to x$ also, whence $y \in \mathcal{C}$. This implies that $P_{|\mathcal{C}}$ is a transition matrix, which clearly must be irreducible on $\mathcal{C}$. Therefore, there exists a unique stationary distribution $\pi^{\mathcal{C}}$ for $P_{|\mathcal{C}}$. Let π be a probability on Ω with $\pi = \pi P$. By Proposition 1.25, $\pi(y) = 0$ for

$y \notin \mathcal{C}$, whence π is supported on $\mathcal{C}$. Consequently, for $x \in \mathcal{C}$,

$$\pi(x) = \sum_{y \in \Omega} \pi(y) P(y,x) = \sum_{y \in \mathcal{C}} \pi(y) P(y,x) = \sum_{y \in \mathcal{C}} \pi(y) P_{|\mathcal{C}}(y,x),$$

and π restricted to $\mathcal{C}$ is stationary for $P_{|\mathcal{C}}$. By uniqueness of the stationary distribution for $P_{|\mathcal{C}}$, it follows that $\pi(x) = \pi^{\mathcal{C}}(x)$ for all $x \in \mathcal{C}$. Therefore,

$$\pi(x) = \begin{cases} \pi^{\mathcal{C}}(x) & \text{if } x \in \mathcal{C}, \\ 0 & \text{if } x \notin \mathcal{C}, \end{cases}$$

and the solution to $\pi = \pi P$ is unique.

Suppose there are distinct essential communicating classes for P, say $\mathcal{C}_1$ and $\mathcal{C}_2$. The restriction of P to each of these classes is irreducible. Thus for $i = 1, 2$, there exists a measure π supported on $\mathcal{C}_i$ which is stationary for $P_{|\mathcal{C}_i}$. Moreover, it is easily verified that each π_i is stationary for P, and so P has more than one stationary distribution. ■

Exercises

EXERCISE 1.1. Let P be the transition matrix of random walk on the n-cycle, where n is odd. Find the smallest value of t such that $P^t(x,y) > 0$ for all states x and y.

EXERCISE 1.2. A graph G is ***connected*** when, for two vertices x and y of G, there exists a sequence of vertices $x_0, x_1, \ldots, x_k$ such that $x_0 = x$, $x_k = y$, and $x_i \sim x_{i+1}$ for $0 \leq i \leq k-1$. Show that random walk on G is irreducible if and only if G is connected.

EXERCISE 1.3. We define a graph to be a ***tree*** if it is connected but contains no cycles. Prove that the following statements about a graph T with n vertices and m edges are equivalent:

(a) T is a tree.
(b) T is connected and $m = n - 1$.
(c) T has no cycles and $m = n - 1$.

EXERCISE 1.4. Let T be a tree. A ***leaf*** is a vertex of degree 1.

(a) Prove that T contains a leaf.
(b) Prove that between any two vertices in T there is a unique simple path.
(c) Prove that T has at least 2 leaves.

EXERCISE 1.5. Let T be a tree. Show that the graph whose vertices are proper 3-colorings of T and whose edges are pairs of colorings which differ at only a single vertex is connected.

EXERCISE 1.6. Let P be an irreducible transition matrix of period b. Show that Ω can be partitioned into b sets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_b$ in such a way that $P(x,y) > 0$ only if $x \in \mathcal{C}_i$ and $y \in \mathcal{C}_{i+1}$. (The addition $i + 1$ is modulo b.)

EXERCISE 1.7. A transition matrix P is ***symmetric*** if $P(x,y) = P(y,x)$ for all $x, y \in \Omega$. Show that if P is symmetric, then the uniform distribution on Ω is stationary for P.

EXERCISE 1.8. Let P be a transition matrix which is reversible with respect to the probability distribution π on Ω. Show that the transition matrix P^2 corresponding to two steps of the chain is also reversible with respect to π.

EXERCISE 1.9. Let π be a stationary distribution for an irreducible transition matrix P. Prove that $\pi(x) > 0$ for all $x \in \Omega$, without using the explicit formula (1.25).

EXERCISE 1.10. Check carefully that equation (1.19) is true.

EXERCISE 1.11. Here we outline another proof, more analytic, of the existence of stationary distributions. Let P be the transition matrix of a Markov chain on a finite state space Ω. For an arbitrary initial distribution μ on Ω and $n > 0$, define the distribution ν_n by

$$\nu_n = \frac{1}{n}\left(\mu + \mu P + \cdots + \mu P^{n-1}\right).$$

(a) Show that for any $x \in \Omega$ and $n > 0$,

$$|\nu_n P(x) - \nu_n(x)| \le \frac{2}{n}.$$

(b) Show that there exists a subsequence $(\nu_{n_k})_{k\ge 0}$ such that $\lim_{k\to\infty} \nu_{n_k}(x)$ exists for every $x \in \Omega$.

(c) For $x \in \Omega$, define $\nu(x) = \lim_{k\to\infty} \nu_{n_k}(x)$. Show that ν is a stationary distribution for P.

EXERCISE 1.12. Let P be the transition matrix of an irreducible Markov chain with state space Ω. Let $B \subset \Omega$ be a non-empty subset of the state space, and assume $h : \Omega \to \mathbb{R}$ is a function harmonic at all states $x \notin B$.

Prove that if h is non-constant and $h(y) = \max_{x\in\Omega} h(x)$, then $y \in B$.

(This is a discrete version of the ***maximum principle***.)

EXERCISE 1.13. Give a direct proof that the stationary distribution for an irreducible chain is unique.

Hint: Given stationary distributions π_1 and π_2, consider the state x that minimizes $\pi_1(x)/\pi_2(x)$ and show that all y with $P(x,y) > 0$ have $\pi_1(y)/\pi_2(y) = \pi_1(x)/\pi_2(x)$.

EXERCISE 1.14. Show that any stationary measure π of an irreducible chain must be strictly positive.

Hint: Show that if $\pi(x) = 0$, then $\pi(y) = 0$ whenever $P(x,y) > 0$.

EXERCISE 1.15. For a subset $A \subset \Omega$, define $f(x) = \mathbf{E}_x(\tau_A)$. Show that

(a)
$$f(x) = 0 \quad \text{for } x \in A. \tag{1.35}$$

(b)
$$f(x) = 1 + \sum_{y\in\Omega} P(x,y)f(y) \quad \text{for } x \notin A. \tag{1.36}$$

(c) f is uniquely determined by (1.35) and (1.36).

The following exercises concern the material in Section 1.7.

EXERCISE 1.16. Show that $\leftrightarrow$ is an equivalence relation on Ω.

EXERCISE 1.17. Show that the set of stationary measures for a transition matrix forms a polyhedron with one vertex for each essential communicating class.

Notes

Markov first studied the stochastic processes that came to be named after him in Markov (1906). See Basharin, Langville, and Naumov (2004) for the early history of Markov chains.

The right-hand side of (1.1) does not depend on t. We take this as part of the definition of a Markov chain; note that other authors sometimes regard this as a special case, which they call ***time homogeneous***. (This simply means that the transition matrix is the same at each step of the chain. It is possible to give a more general definition in which the transition matrix depends on t. We will not consider such chains in this book.)

Aldous and Fill (1999, Chapter 2, Proposition 4) present a version of the key computation for Proposition 1.14 which requires only that the initial distribution of the chain equals the distribution of the chain when it stops. We have essentially followed their proof.

The standard approach to demonstrating that irreducible aperiodic Markov chains have unique stationary distributions is through the Perron-Frobenius theorem. See, for instance, Karlin and Taylor (1975) or Seneta (2006).

See Feller (1968, Chapter XV) for the classification of states of Markov chains.

Complements. The following lemma is needed for the proof of Proposition 1.7. We include a proof here for completeness.

LEMMA 1.27. *If $S \subset \mathbb{Z}^+$ has $\gcd(S) = g_S$, then there is some integer m_S such that for all $m \geq m_S$ the product mg_S can be written as a linear combination of elements of S with non-negative integer coefficients.*

PROOF. *Step 1.* Given $S \subset \mathbb{Z}^+$ nonempty, define $g_S^\star$ as the smallest positive integer which is an integer combination of elements of S (the smallest positive element of the additive group generated by S). Then $g_S^\star$ divides every element of S (otherwise, consider the remainder) and g_S must divide $g_S^\star$, so $g_S^\star = g_S$.

Step 2. For any set S of positive integers, there is a finite subset F such that $\gcd(S) = \gcd(F)$. Indeed the non-increasing sequence $\gcd(S \cap [1, n])$ can strictly decrease only finitely many times, so there is a last time. Thus it suffices to prove the fact for finite subsets F of $\mathbb{Z}^+$; we start with sets of size 2 (size 1 is a tautology) and then prove the general case by induction on the size of F.

Step 3. Let $F = \{a, b\} \subset \mathbb{Z}^+$ have $\gcd(F) = g$. Given $m > 0$, write $mg = ca+db$ for some integers c, d. Observe that c, d are not unique since $mg = (c + kb)a + (d - ka)b$ for any k. Thus we can write $mg = ca + db$ where $0 \leq c < b$. If $mg > (b-1)a - b$, then we must have $d \geq 0$ as well. Thus for $F = \{a, b\}$ we can take $m_F = (ab - a - b)/g + 1$.

Step 4 (The induction step). Let F be a finite subset of $\mathbb{Z}^+$ with $\gcd(F) = g_F$. Then for any $a \in \mathbb{Z}^+$ the definition of gcd yields that $g := \gcd(\{a\} \cup F) = \gcd(a, g_F)$. Suppose that n satisfies $ng \geq m_{\{a,g_F\}}g + m_F g_F$. Then we can write $ng - m_F g_F = ca + dg_F$ for integers $c, d \geq 0$. Therefore $ng = ca + (d + m_F)g_F = ca + \sum_{f \in F} c_f f$ for some integers $c_f \geq 0$ by the definition of m_F. Thus we can take $m_{\{a\} \cup F} = m_{\{a,g_F\}} + m_F g_F / g$. ■

CHAPTER 2

Classical (and Useful) Markov Chains

Here we present several basic and important examples of Markov chains. The results we prove in this chapter will be used in many places throughout the book.

This is also the only chapter in the book where the central chains are not always irreducible. Indeed, two of our examples, gambler's ruin and coupon collecting, both have absorbing states. For each we examine closely how long it takes to be absorbed.

2.1. Gambler's Ruin

Consider a gambler betting on the outcome of a sequence of independent fair coin tosses. If the coin comes up heads, she adds one dollar to her purse; if the coin lands tails up, she loses one dollar. If she ever reaches a fortune of n dollars, she will stop playing. If her purse is ever empty, then she must stop betting.

The gambler's situation can be modeled by a random walk on a path with vertices $\{0, 1, \dots, n\}$. At all interior vertices, the walk is equally likely to go up by 1 or down by 1. That states 0 and n are absorbing, meaning that once the walk arrives at either 0 or n, it stays forever (cf. Section 1.7).

There are two questions that immediately come to mind: how long will it take for the gambler to arrive at one of the two possible fates? What are the probabilities of the two possibilities?

PROPOSITION 2.1. *Assume that a gambler making fair unit bets on coin flips will abandon the game when her fortune falls to* 0 *or rises to* n. *Let* X_t *be gambler's fortune at time* t *and let* τ *be the time required to be absorbed at one of* 0 *or* n. *Assume that* $X_0 = k$, *where* $0 \leq k \leq n$. *Then*

$$\mathbf{P}_k\{X_\tau = n\} = k/n \tag{2.1}$$

and

$$\mathbf{E}_k(\tau) = k(n-k). \tag{2.2}$$

PROOF. Let p_k be the probability that the gambler reaches a fortune of n before ruin, given that she starts with k dollars. We solve simultaneously for $p_0, p_1, \dots, p_n$. Clearly $p_0 = 0$ and $p_n = 1$, while

$$p_k = \frac{1}{2}p_{k-1} + \frac{1}{2}p_{k+1} \quad \text{for } 1 \leq k \leq n-1. \tag{2.3}$$

Why? With probability $1/2$, the walk moves to $k+1$. The conditional probability of reaching n before 0, starting from $k+1$, is exactly p_{k+1}. Similarly, with probability $1/2$ the walk moves to $k-1$, and the conditional probability of reaching n before 0 from state $k-1$ is p_{k-1}.

Solving the system (2.3) of linear equations yields $p_k = k/n$ for $0 \leq k \leq n$.

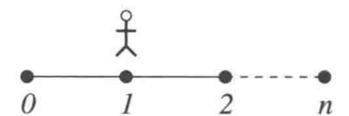


FIGURE 2.1. How long until the walk reaches either 0 or n? What is the probability of each?

For (2.2), again we try to solve for all the values at once. To this end, write f_k for the expected time $\mathbf{E}_k(\tau)$ to be absorbed, starting at position k. Clearly, $f_0 = f_n = 0$; the walk is started at one of the absorbing states. For $1 \leq k \leq n-1$, it is true that

$$f_k = \frac{1}{2}\left(1 + f_{k+1}\right) + \frac{1}{2}\left(1 + f_{k-1}\right). \tag{2.4}$$

Why? When the first step of the walk increases the gambler's fortune, then the conditional expectation of τ is 1 (for the initial step) plus the expected additional time needed. The expected additional time needed is f_{k+1}, because the walk is now at position $k+1$. Parallel reasoning applies when the gambler's fortune first decreases.

Exercise 2.1 asks the reader to solve this system of equations, completing the proof of (2.2). ■

REMARK 2.2. See Chapter 9 for powerful generalizations of the simple methods we have just applied.

2.2. Coupon Collecting

A company issues n different types of coupons. A collector desires a complete set. We suppose each coupon he acquires is equally likely to be each of the n types. How many coupons must he obtain so that his collection contains all n types?

It may not be obvious why this is a Markov chain. Let X_t denote the number of different types represented among the collector's first t coupons. Clearly $X_0 = 0$. When the collector has coupons of k different types, there are $n-k$ types missing. Of the n possibilities for his next coupon, only $n-k$ will expand his collection. Hence

$$\mathbf{P}\{X_{t+1} = k+1 \mid X_t = k\} = \frac{n-k}{n}$$

and

$$\mathbf{P}\{X_{t+1} = k \mid X_t = k\} = \frac{k}{n}.$$

Every trajectory of this chain is non-decreasing. Once the chain arrives at state n (corresponding to a complete collection), it is absorbed there. We are interested in the number of steps required to reach the absorbing state.

PROPOSITION 2.3. *Consider a collector attempting to collect a complete set of coupons. Assume that each new coupon is chosen uniformly and independently from the set of n possible types, and let τ be the (random) number of coupons collected when the set first contains every type. Then*

$$\mathbf{E}(\tau) = n \sum_{k=1}^{n} \frac{1}{k}.$$

PROOF. The expectation $\mathbf{E}(\tau)$ can be computed by writing τ as a sum of geometric random variables. Let τ_k be the total number of coupons accumulated when the collection first contains k distinct coupons. Then

$$\tau = \tau_n = \tau_1 + (\tau_2 - \tau_1) + \cdots + (\tau_n - \tau_{n-1}). \tag{2.5}$$

Furthermore, $\tau_k - \tau_{k-1}$ is a geometric random variable with success probability $(n-k+1)/n$: after collecting τ_{k-1} coupons, there are $n-k+1$ types missing from the collection. Each subsequent coupon drawn has the same probability $(n-k+1)/n$ of being a type not already collected, until a new type is finally drawn. Thus $\mathbf{E}(\tau_k - \tau_{k-1}) = n/(n-k+1)$ and

$$\mathbf{E}(\tau) = \sum_{k=1}^{n} \mathbf{E}(\tau_k - \tau_{k-1}) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k}. \tag{2.6}$$

■

While the argument for Proposition 2.3 is simple and vivid, we will often need to know more about the distribution of τ in future applications. Recall that $|\sum_{k=1}^{n} 1/k - \log n| \leq 1$, whence $|\mathbf{E}(\tau) - n\log n| \leq n$ (see Exercise 2.4 for a better estimate). Proposition 2.4 says that τ is unlikely to be much larger than its expected value.

PROPOSITION 2.4. *Let τ be a coupon collector random variable, as in Proposition 2.3. For any $c > 0$,*

$$\mathbf{P}\{\tau > \lceil n\log n + cn \rceil\} \leq e^{-c}. \tag{2.7}$$

PROOF. Let A_i be the event that the i-th type does not appear among the first $\lceil n\log n + cn \rceil$ coupons drawn. Observe first that

$$\mathbf{P}\{\tau > \lceil n\log n + cn \rceil\} = \mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbf{P}(A_i).$$

Since each trial has probability $1 - n^{-1}$ of *not* drawing coupon i and the trials are independent, the right-hand side above is bounded above by

$$\sum_{i=1}^{n} \left(1 - \frac{1}{n}\right)^{\lceil n\log n + cn \rceil} \leq n \exp\left(-\frac{n\log n + cn}{n}\right) = e^{-c},$$

proving (2.7). ■

2.3. The Hypercube and the Ehrenfest Urn Model

The ***n*-dimensional hypercube** is a graph whose vertices are the binary n-tuples $\{0,1\}^n$. Two vertices are connected by an edge when they differ in exactly one coordinate. See Figure 2.2 for an illustration of the three-dimensional hypercube.

The simple random walk on the hypercube moves from a vertex $(x^1, x^2, \ldots, x^n)$ by choosing a coordinate $j \in \{1, 2, \ldots, n\}$ uniformly at random and setting the new state equal to $(x^1, \ldots, x^{j-1}, 1 - x^j, x^{j+1}, \ldots, x^n)$. That is, the bit at the walk's chosen coordinate is flipped. (This is a special case of the walk defined in Section 1.4.)

Unfortunately, the simple random walk on the hypercube is periodic, since every move flips the parity of the number of 1's. The ***lazy random walk***, which does not have this problem, remains at its current position with probability 1/2 and moves

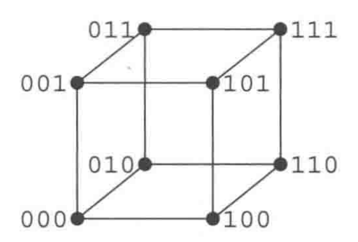


FIGURE 2.2. The three-dimensional hypercube.

as above with probability 1/2. This chain can be realized by choosing a coordinate uniformly at random and *refreshing* the bit at this coordinate by replacing it with an unbiased random bit independent of time, current state, and coordinate chosen.

Since the hypercube is an n-regular graph, Example 1.12 implies that the stationary distribution of both the simple and lazy random walks is uniform on $\{0,1\}^n$.

We now consider a process, the ***Ehrenfest urn***, which at first glance appears quite different. Suppose n balls are distributed among two urns, I and II. At each move, a ball is selected uniformly at random and transferred from its current urn to the other urn. If X_t is the number of balls in urn I at time t, then the transition matrix for (X_t) is

$$P(j,k) = \begin{cases} \frac{n-j}{n} & \text{if } k = j+1, \\ \frac{j}{n} & \text{if } k = j-1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Thus (X_t) is a Markov chain with state space $\Omega = \{0,1,2,\dots,n\}$ that moves by ± 1 on each move and is biased towards the middle of the interval. The stationary distribution for this chain is binomial with parameters n and $1/2$ (see Exercise 2.5).

The Ehrenfest urn is a projection (in a sense that will be defined precisely in Section 2.3.1) of the random walk on the n-dimensional hypercube. This is unsurprising given the standard bijection between $\{0,1\}^n$ and subsets of $\{1,\dots,n\}$, under which a set corresponds to the vector with 1's in the positions of its elements. We can view the position of the random walk on the hypercube as specifying the set of balls in Ehrenfest urn I; then changing a bit corresponds to moving a ball into or out of the urn.

Define the ***Hamming weight*** $W(\boldsymbol{x})$ of a vector $\boldsymbol{x} := (x^1,\dots,x^n) \in \{0,1\}^n$ to be its number of coordinates with value 1:

$$W(\boldsymbol{x}) = \sum_{j=1}^{n} x^j. \tag{2.9}$$

Let $(\boldsymbol{X}_t)$ be the simple random walk on the n-dimensional hypercube, and let $W_t = W(\boldsymbol{X}_t)$ be the Hamming weight of the walk's position at time t.

When $W_t = j$, the weight increments by a unit amount when one of the $n-j$ coordinates with value 0 is selected. Likewise, when one of the j coordinates with value 1 is selected, the weight decrements by one unit. From this description, it is clear that (W_t) is a Markov chain with transition probabilities given by (2.8).

2.3.1. Projections of chains. The Ehrenfest urn is a *projection*, which we define in this section, of the simple random walk on the hypercube.

Assume that we are given a Markov chain $(X_0, X_1, \dots)$ with state space Ω and transition matrix P and also some equivalence relation that partitions Ω into equivalence classes. We denote the equivalence class of $x \in \Omega$ by $[x]$. (For the Ehrenfest example, two bitstrings are equivalent when they contain the same number of 1's.)

Under what circumstances will $([X_0], [X_1], \dots)$ also be a Markov chain? For this to happen, knowledge of what equivalence class we are in at time t must suffice to determine the distribution over equivalence classes at time $t+1$. If the probability $P(x, [y])$ is always the same as $P(x', [y])$ when x and x' are in the same equivalence class, that is clearly enough. We summarize this in the following lemma.

LEMMA 2.5. *Let Ω be the state space of a Markov chain (X_t) with transition matrix P. Let $\sim$ be an equivalence relation on Ω with equivalence classes $\Omega^\sharp = \{[x] : x \in \Omega\}$, and assume that P satisfies*

$$P(x, [y]) = P(x', [y]) \tag{2.10}$$

whenever $x \sim x'$. Then $[X_t]$ is a Markov chain with state space $\Omega^\sharp$ and transition matrix $P^\sharp$ defined by $P^\sharp([x], [y]) := P(x, [y])$.

The process of constructing a new chain by taking equivalence classes for an equivalence relation compatible with the transition matrix (in the sense of (2.10)) is called ***projection***, or sometimes ***lumping***.

2.4. The Pólya Urn Model

Consider the following process, known as ***Pólya's urn***. Start with an urn containing two balls, one black and one white. From this point on, proceed by choosing a ball at random from those already in the urn; return the chosen ball to the urn and add another ball of the same color. If there are j black balls in the urn after k balls have been added (so that there are $k + 2$ balls total in the urn), then the probability that another black ball is added is $j/(k + 2)$. The sequence of ordered pairs listing the numbers of black and white balls is a Markov chain with state space $\{1, 2, \dots\}^2$.

LEMMA 2.6. *Let B_k be the number of black balls in Pólya's urn after the addition of k balls. The distribution of B_k is uniform on $\{1, 2, \dots, k + 1\}$.*

PROOF. Let $U_0, U_1, \dots, U_n$ be independent and identically distributed random variables, each uniformly distributed on the interval $[0, 1]$. Let

$$L_k := |\{j \in \{0, 1, \dots, k\} : U_j \le U_0\}|$$

be the number of $U_0, U_1, \dots, U_k$ which are less than or equal to U_0.

The event $\{L_k = j,\ L_{k+1} = j + 1\}$ occurs if and only if U_0 is the $(j + 1)$-st smallest and U_{k+1} is one of the $j + 1$ smallest among $\{U_0, U_1, \dots, U_{k+1}\}$. There are $j(k!)$ orderings of $\{U_0, U_1, \dots, U_{k+1}\}$ making up this event; since all $(k + 2)!$ orderings are equally likely,

$$\mathbf{P}\{L_k = j,\ L_{k+1} = j + 1\} = \frac{j(k!)}{(k + 2)!} = \frac{j}{(k + 2)(k + 1)}. \tag{2.11}$$

Since each relative ordering of $U_0, \dots, U_k$ is equally likely, we have $\mathbf{P}\{L_k = j\} = 1/(k + 1)$. Together with (2.11) this implies that

$$\mathbf{P}\{L_{k+1} = j + 1 \mid L_k = j\} = \frac{j}{k + 2}. \tag{2.12}$$

Since $L_{k+1} \in \{j, j+1\}$ given $L_k = j$,

$$\mathbf{P}\{L_{k+1} = j \mid L_k = j\} = \frac{k+2-j}{k+2}. \tag{2.13}$$

Note that L_1 and B_1 have the same distribution. By (2.12) and (2.13), the sequences $(L_k)_{k=1}^n$ and $(B_k)_{k=1}^n$ have the same transition probabilities. Hence the sequences $(L_k)_{k=1}^n$ and $(B_k)_{k=1}^n$ have the same distribution. In particular, L_k and B_k have the same distribution.

Since the position of U_0 among $\{U_0, \ldots, U_k\}$ is uniform among the $k+1$ possible positions, it follows that L_k is uniform on $\{1, \ldots, k+1\}$. Thus, B_k is uniform on $\{1, \ldots, k+1\}$. ■

REMARK 2.7. Lemma 2.6 can also be proved by showing that $\mathbf{P}\{B_k = j\} = 1/(k+1)$ for all $j = 1, \ldots, k+1$ using induction on k.

2.5. Birth-and-Death Chains

A ***birth-and-death chain*** has state space $\Omega = \{0, 1, 2, \ldots, n\}$. In one step the state can increase or decrease by at most 1. The current state can be thought of as the size of some population; in a single step of the chain there can be at most one birth or death. The transition probabilities can be specified by $\{(p_k, r_k, q_k)\}_{k=0}^n$, where $p_k + r_k + q_k = 1$ for each k and

- p_k is the probability of moving from k to $k+1$ when $0 \le k < n$,
- q_k is the probability of moving from k to $k-1$ when $0 < k \le n$,
- r_k is the probability of remaining at k when $0 \le k \le n$,
- $q_0 = p_n = 0$.

PROPOSITION 2.8. *Every birth-and-death chain is reversible.*

PROOF. A function w on Ω satisfies the detailed balance equations (1.30) if and only if

$$p_{k-1} w_{k-1} = q_k w_k$$

for $1 \le k \le n$. For our birth-and-death chain, a solution is given by $w_0 = 1$ and

$$w_k = \prod_{i=1}^{k} \frac{p_{i-1}}{q_i}$$

for $1 \le k \le n$. Normalizing so that the sum is unity yields

$$\pi_k = \frac{w_k}{\sum_{j=0}^n w_j}$$

for $0 \le k \le n$. (By Proposition 1.19, π is also a stationary distribution.) ■

Now, fix $\ell \in \{0, 1, \ldots, n\}$. Consider restricting the original chain to $\{0, 1, \ldots, \ell\}$:

- For any $k \in \{0, 1, \ldots, \ell - 1\}$, the chain makes transitions from k as before, moving down with probability q_k, remaining in place with probability r_k, and moving up with probability p_k.
- At ℓ, the chain either moves down or remains in place, with probabilities q_ℓ and $r_\ell + p_\ell$, respectively.

We write $\tilde{\mathbf{E}}$ for expectations for this new chain. By the proof of Proposition 2.8, the stationary probability $\tilde{\pi}$ of the truncated chain is given by

$$\tilde{\pi}_k = \frac{w_k}{\sum_{j=0}^{\ell} w_j}$$

for $0 \leq k \leq \ell$. Since in the truncated chain the only possible moves from ℓ are to stay put or to step down to $\ell - 1$, the expected first return time $\tilde{\mathbf{E}}_\ell(\tau_\ell^+)$ satisfies

$$\tilde{\mathbf{E}}_\ell(\tau_\ell^+) = (r_\ell + p_\ell) \cdot 1 + q_\ell \left(\tilde{\mathbf{E}}_{\ell-1}(\tau_\ell) + 1 \right) = 1 + q_\ell \tilde{\mathbf{E}}_{\ell-1}(\tau_\ell). \tag{2.14}$$

By Proposition 1.14(ii),

$$\tilde{\mathbf{E}}_\ell(\tau_\ell^+) = \frac{1}{\tilde{\pi}(\ell)} = \frac{1}{w_\ell} \sum_{j=0}^{\ell} w_j. \tag{2.15}$$

We have constructed the truncated chain so that $\tilde{\mathbf{E}}_{\ell-1}(\tau_\ell) = \mathbf{E}_{\ell-1}(\tau_\ell)$. Rearranging (2.14) and (2.15) gives

$$\mathbf{E}_{\ell-1}(\tau_\ell) = \frac{1}{q_\ell} \left(\sum_{j=0}^{\ell} \frac{w_j}{w_\ell} - 1 \right) = \frac{1}{q_\ell w_\ell} \sum_{j=0}^{l-1} w_j. \tag{2.16}$$

To find $\mathbf{E}_a(\tau_b)$ for $a < b$, just sum:

$$\mathbf{E}_a(\tau_b) = \sum_{\ell=a+1}^{b} \mathbf{E}_{\ell-1}(\tau_\ell).$$

Consider two important special cases. Suppose that

$$(p_k, r_k, q_k) = (p, r, q) \text{ for } 1 \leq k < n,$$
$$(p_0, r_0, q_0) = (p, r+q, 0), \quad (p_n, r_n, q_n) = (0, r+p, q)$$

for $p, r, q \geq 0$ with $p + r + q = 1$. First consider the case where $p \neq q$. We have $w_k = (p/q)^k$ for $0 \leq k \leq n$, and from (2.16), for $1 \leq \ell \leq n$,

$$\mathbf{E}_{\ell-1}(\tau_\ell) = \frac{1}{q(p/q)^\ell} \sum_{j=0}^{\ell-1} (p/q)^j = \frac{(p/q)^\ell - 1}{q(p/q)^\ell [(p/q) - 1]} = \frac{1}{p-q} \left[1 - \left(\frac{q}{p} \right)^\ell \right].$$

If $p = q$, then $w_j = 1$ for all j and

$$\mathbf{E}_{\ell-1}(\tau_\ell) = \frac{\ell}{p}.$$

2.6. Random Walks on Groups

Several of the examples we have already examined and many others we will study in future chapters share important symmetry properties, which we make explicit here. Recall that a ***group*** is a set G endowed with an associative operation $\cdot : G \times G \to G$ and an ***identity*** $\text{id} \in G$ such that for all $g \in G$,

(i) $\text{id} \cdot g = g$ and $g \cdot \text{id} = g$.
(ii) there exists an ***inverse*** $g^{-1} \in G$ for which $g \cdot g^{-1} = g^{-1} \cdot g = \text{id}$.

Given a probability distribution μ on a group $(G, \cdot)$, we define the ***random walk on G with increment distribution*** μ as follows: it is a Markov chain with state space G and which moves by multiplying the current state *on the left* by a random element of G selected according to μ. Equivalently, the transition matrix P of this chain has entries

$$P(g, hg) = \mu(h)$$

for all $g, h \in G$.

REMARK 2.9. We multiply the current state by the increment *on the left* because it is generally more natural in non-commutative examples, such as the symmetric group—see Section 8.1.3. For commutative examples, such as the two described immediately below, it of course does not matter on which side we multiply.

EXAMPLE 2.10 (The n-cycle). Let μ assign probability $1/2$ to each of 1 and $n-1 \equiv -1 \pmod n$ in the additive cyclic group $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$. The ***simple random walk on the n-cycle*** first introduced in Example 1.4 is the random walk on $\mathbb{Z}_n$ with increment distribution μ. Similarly, let ν assign weight $1/4$ to both 1 and $n-1$ and weight $1/2$ to 0. Then ***lazy random walk on the n-cycle***, discussed in Example 1.8, is the random walk on $\mathbb{Z}_n$ with increment distribution ν.

EXAMPLE 2.11 (The hypercube). The hypercube random walks defined in Section 2.3 are random walks on the group $\mathbb{Z}_2^n$, which is the direct product of n copies of the two-element group $\mathbb{Z}_2 = \{0, 1\}$. For the simple random walk the increment distribution is uniform on the set $\{\mathbf{e}_i : 1 \le i \le n\}$, where the vector $\mathbf{e}_i$ has a 1 in the i-th place and 0 in all other entries. For the lazy version, the increment distribution gives the vector $\mathbf{0}$ (with all zero entries) weight $1/2$ and each $\mathbf{e}_i$ weight $1/2n$.

PROPOSITION 2.12. *Let P be the transition matrix of a random walk on a finite group G and let U be the uniform probability distribution on G. Then U is a stationary distribution for P.*

PROOF. Let μ be the increment distribution of the random walk. For any $g \in G$,

$$\sum_{h \in G} U(h)P(h, g) = \frac{1}{|G|} \sum_{k \in G} P(k^{-1}g, g) = \frac{1}{|G|} \sum_{k \in G} \mu(k) = \frac{1}{|G|} = U(g).$$

For the first equality, we re-indexed by setting $k = gh^{-1}$. ■

2.6.1. Generating sets, irreducibility, Cayley graphs, and reversibility. For a set $H \subset G$, let $\langle H \rangle$ be the smallest group containing all the elements of H; recall that every element of $\langle H \rangle$ can be written as a product of elements in H and their inverses. A set H is said to ***generate*** G if $\langle H \rangle = G$.

PROPOSITION 2.13. *Let μ be a probability distribution on a finite group G. The random walk on G with increment distribution μ is irreducible if and only if $S = \{g \in G : \mu(g) > 0\}$ generates G.*

PROOF. Let a be an arbitrary element of G. If the random walk is irreducible, then there exists an $r > 0$ such that $P^r(\mathrm{id}, a) > 0$. In order for this to occur, there must be a sequence $s_1, \ldots, s_r \in G$ such that $a = s_r s_{r-1} \ldots s_1$ and $s_i \in S$ for $i = 1, \ldots, r$. Thus $a \in \langle S \rangle$.

Now assume S generates G, and consider $a, b \in G$. We know that ba^{-1} can be written as a word in the elements of S and their inverses. Since every element of G

has finite order, any inverse appearing in the expression for ba^{-1} can be rewritten as a positive power of the same group element. Let the resulting expression be $ba^{-1} = s_r s_{r-1} \dots s_1$, where $s_i \in S$ for $i = 1, \dots, r$. Then

$$\begin{aligned} P^m(a,b) &\geq P(a, s_1 a) P(s_1 a, s_2 s_1 a) \cdots P(s_{r-1} s_{r-2} \dots s_1 a, (ba^{-1})a) \\ &= \mu(s_1)\mu(s_2)\dots\mu(s_r) > 0. \end{aligned}$$

■

When S is a set which generates a finite group G, the ***directed Cayley graph*** associated to G and S is the directed graph with vertex set G in which (v, w) is an edge if and only if $v = sw$ for some generator $s \in S$.

We call a set S of generators of G ***symmetric*** if $s \in S$ implies $s^{-1} \in S$. When S is symmetric, all edges in the directed Cayley graph are bidirectional, and it may be viewed as an ordinary graph. When G is finite and S is a symmetric set that generates G, the simple random walk (as defined in Section 1.4) on the corresponding Cayley graph is the same as the random walk on G with increment distribution μ taken to be the uniform distribution on S.

In parallel fashion, we call a probability distribution μ on a group G ***symmetric*** if $\mu(g) = \mu(g^{-1})$ for every $g \in G$.

PROPOSITION 2.14. *The random walk on a finite group G with increment distribution μ is reversible if μ is symmetric.*

PROOF. Let U be the uniform probability distribution on G. For any $g, h \in G$, we have that

$$U(g)P(g,h) = \frac{\mu(hg^{-1})}{|G|} \quad \text{and} \quad U(h)P(h,g) = \frac{\mu(gh^{-1})}{|G|}$$

are equal if and only if $\mu(hg^{-1}) = \mu((hg^{-1})^{-1})$. ■

REMARK 2.15. The converse of Proposition 2.14 is also true; see Exercise 2.7.

2.6.2. Transitive chains. A Markov chain is called ***transitive*** if for each pair $(x, y) \in \Omega \times \Omega$ there is a bijection $\varphi = \varphi_{(x,y)} : \Omega \to \Omega$ such that

$$\varphi(x) = y \quad \text{and} \quad P(z, w) = P(\varphi(z), \varphi(w)) \text{ for all } z, w \in \Omega. \tag{2.17}$$

Roughly, this means the chain "looks the same" from any point in the state space Ω. Clearly any random walk on a group is transitive; set $\varphi_{(x,y)}(g) = gx^{-1}y$. However, there are examples of transitive chains that are not random walks on groups; see McKay and Praeger (1996).

Many properties of random walks on groups generalize to the transitive case, including Proposition 2.12.

PROPOSITION 2.16. *Let P be the transition matrix of a transitive Markov chain on a finite state space Ω. Then the uniform probability distribution on Ω is stationary for P.*

PROOF. Fix $x, y \in \Omega$ and let $\varphi : \Omega \to \Omega$ be a transition-probability-preserving bijection for which $\varphi(x) = y$. Let U be the uniform probability on Ω. Then

$$\sum_{z \in \Omega} U(z)P(z,x) = \sum_{z \in \Omega} U(\varphi(z))P(\varphi(z), y) = \sum_{w \in \Omega} U(w)P(w,y),$$

where we have re-indexed with $w = \varphi(z)$. We have shown that when the chain is started in the uniform distribution and run one step, the total weight arriving at each state is the same. Since $\sum_{x,z\in\Omega} U(z)P(z,x) = 1$, we must have

$$\sum_{z\in\Omega} U(z)P(z,x) = \frac{1}{|\Omega|} = U(x).$$

■

2.7. Random Walks on $\mathbb{Z}$ and Reflection Principles

A ***nearest-neighbor random walk*** on $\mathbb{Z}$ moves right and left by at most one step on each move, and each move is independent of the past. More precisely, if (Δ_t) is a sequence of independent and identically distributed $\{-1,0,1\}$-valued random variables and $X_t = \sum_{s=1}^{t} \Delta_s$, then the sequence (X_t) is a nearest-neighbor random walk with increments (Δ_t).

This sequence of random variables is a Markov chain with infinite state space $\mathbb{Z}$ and transition matrix

$$P(k,k+1) = p, \quad P(k,k) = r, \quad P(k,k-1) = q,$$

where $p + r + q = 1$.

The special case where $p = q = 1/2$, $r = 0$ is the simple random walk on $\mathbb{Z}$, as defined in Section 1.4. In this case

$$\mathbf{P}_0\{X_t = k\} = \begin{cases} \binom{t}{\frac{t-k}{2}} 2^{-t} & \text{if } t-k \text{ is even,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.18}$$

since there are $\binom{t}{\frac{t-k}{2}}$ possible paths of length t from 0 to k.

When $p = q = 1/4$ and $r = 1/2$, the chain is the lazy simple random walk on $\mathbb{Z}$. (Recall the definition of lazy chains in Section 1.3.)

THEOREM 2.17. *Let (X_t) be simple random walk on $\mathbb{Z}$, and recall that*

$$\tau_0 = \min\{t \geq 0 : X_t = 0\}$$

is the first time the walk hits zero. Then

$$\mathbf{P}_k\{\tau_0 > r\} \leq \frac{12k}{\sqrt{r}} \tag{2.19}$$

for any integers $k, r > 0$.

We prove this by a sequence of lemmas which are of independent interest.

LEMMA 2.18 (Reflection Principle). *Let (X_t) be either the simple random walk or the lazy simple random walk on $\mathbb{Z}$. For any positive integers j, k, and r,*

$$\mathbf{P}_k\{\tau_0 < r, X_r = j\} = \mathbf{P}_k\{X_r = -j\} \tag{2.20}$$

and

$$\mathbf{P}_k\{\tau_0 < r, X_r > 0\} = \mathbf{P}_k\{X_r < 0\}. \tag{2.21}$$

PROOF. By the Markov property, the walk "starts afresh" from 0 when it hits 0, meaning that the walk viewed from the first time it hits zero is independent of its past and has the same distribution as a walk started from zero. Hence for any $s < r$ and $j > 0$ we have

$$\mathbf{P}_k\{\tau_0 = s, X_r = j\} = \mathbf{P}_k\{\tau_0 = s\}\mathbf{P}_0\{X_{r-s} = j\}.$$

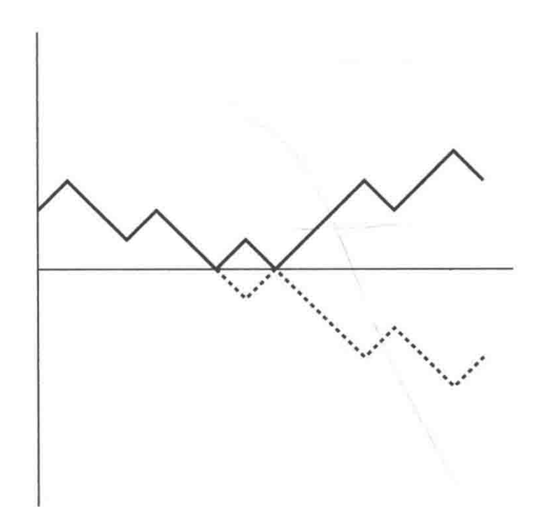


FIGURE 2.3. A path hitting zero and ending above zero can be transformed, by reflection, into a path ending below zero.

The distribution of X_t is symmetric when started at 0, so the right-hand side is equal to

$$\mathbf{P}_k\{\tau_0 = s\}\mathbf{P}_0\{X_{r-s} = -j\} = \mathbf{P}_k\{\tau_0 = s, X_r = -j\}.$$

Summing over $s < r$, we obtain

$$\mathbf{P}_k\{\tau_0 < r, X_r = j\} = \mathbf{P}_k\{\tau_0 < r, X_r = -j\} = \mathbf{P}_k\{X_r = -j\}.$$

To justify the last equality, note that a random walk started from $k > 0$ must pass through 0 before reaching a negative integer.

Finally, summing (2.20) over all $j > 0$ yields (2.21). ∎

REMARK 2.19. There is also a simple combinatorial interpretation of the proof of Lemma 2.18. There is a one-to-one correspondence between walk paths which hit 0 before time r and are positive at time r and walk paths which are negative at time r. This is illustrated in Figure 2.3: to obtain a bijection from the former set of paths to the latter set, reflect a path after the first time it hits 0.

EXAMPLE 2.20 (First passage time for simple random walk). A nice application of Lemma 2.18 gives the distribution of τ_0 when starting from 1 for simple random walk on $\mathbb{Z}$. We have

$$\begin{aligned}\mathbf{P}_1\{\tau_0 = 2m+1\} &= \mathbf{P}_1\{\tau_0 > 2m, X_{2m} = 1, X_{2m+1} = 0\} \\ &= \mathbf{P}_1\{\tau_0 > 2m, X_{2m} = 1\} \cdot \mathbf{P}_1\{X_{2m+1} = 0 \mid X_{2m} = 1\} \\ &= \mathbf{P}_1\{\tau_0 > 2m, X_{2m} = 1\} \cdot \left(\frac{1}{2}\right).\end{aligned}$$

Rewriting and using Lemma 2.18 yields

$$\begin{aligned}\mathbf{P}_1\{\tau_0 = 2m+1\} &= \frac{1}{2}\Big[\mathbf{P}_1\{X_{2m} = 1\} - \mathbf{P}_1\{\tau_0 \le 2m, X_{2m} = 1\}\Big] \\ &= \frac{1}{2}\Big[\mathbf{P}_1\{X_{2m} = 1\} - \mathbf{P}_1\{X_{2m} = -1\}\Big].\end{aligned}$$

Substituting using (2.18) shows that

$$\mathbf{P}_1\{\tau_0 = 2m+1\} = \frac{1}{2}\left[\binom{2m}{m}2^{-2m} - \binom{2m}{m-1}2^{-2m}\right] = \frac{1}{(m+1)2^{2m+1}}\binom{2m}{m}.$$

The right-hand side above equals $C_m/2^{2m+1}$, where C_m is the m-th ***Catalan number***.

LEMMA 2.21. *When (X_t) is simple random walk or lazy simple random walk on $\mathbb{Z}$, we have*

$$\mathbf{P}_k\{\tau_0 > r\} = \mathbf{P}_0\{-k < X_r \leq k\}$$

for any $k > 0$.

PROOF. Observe that

$$\mathbf{P}_k\{X_r > 0\} = \mathbf{P}_k\{X_r > 0, \tau_0 \leq r\} + \mathbf{P}_k\{\tau_0 > r\}.$$

By Lemma 2.18,

$$\mathbf{P}_k\{X_r > 0\} = \mathbf{P}_k\{X_r < 0\} + \mathbf{P}_k\{\tau_0 > r\}.$$

By symmetry of the walk, $\mathbf{P}_k\{X_r < 0\} = \mathbf{P}_k\{X_r > 2k\}$, and so

$$\begin{aligned}\mathbf{P}_k\{\tau_0 > r\} &= \mathbf{P}_k\{X_r > 0\} - \mathbf{P}_k\{X_r > 2k\} \\ &= \mathbf{P}_k\{0 < X_r \leq 2k\} = \mathbf{P}_0\{-k < X_r \leq k\}.\end{aligned}$$

■

LEMMA 2.22. *For the simple random walk (X_t) on $\mathbb{Z}$,*

$$\mathbf{P}_0\{X_t = k\} \leq \frac{3}{\sqrt{t}}. \tag{2.22}$$

REMARK 2.23. By applying Stirling's formula a bit more carefully than we do in the proof below, one can see that in fact

$$\mathbf{P}_0\{X_{2r} = 2k\} \leq \frac{1}{\sqrt{\pi r}}\left[1 + o(1)\right].$$

Hence the constant 3 is nowhere near the best possible. Our goal here is to give an explicit upper bound valid for all k without working too hard to achieve the best possible constant. Indeed, note that for simple random walk, if t and k have different parities, the probability on the left-hand side of (2.22) is 0.

PROOF. If $X_{2r} = 2k$, there are $r+k$ "up" moves and $r-k$ "down" moves. The probability of this is $\binom{2r}{r+k}2^{-2r}$. The reader should check that $\binom{2r}{r+k}$ is maximized at $k = 0$, so for $k = 0, 1, \ldots, r$,

$$\mathbf{P}_0\{X_{2r} = 2k\} \leq \binom{2r}{r}2^{-2r} = \frac{(2r)!}{(r!)^2 2^{2r}}.$$

By Stirling's formula (use the bounds $1 \leq e^{1/(12n+1)} \leq e^{1/(12n)} \leq 2$ in (A.10)), we obtain the bound

$$\mathbf{P}_0\{X_{2r} = 2k\} \leq \sqrt{\frac{8}{\pi}}\frac{1}{\sqrt{2r}}. \tag{2.23}$$

To bound $\mathbf{P}_0\{X_{2r+1} = 2k+1\}$, condition on the first step of the walk and use the bound above. Then use the simple bound $[t/(t-1)]^{1/2} \leq \sqrt{2}$ to see that

$$\mathbf{P}_0\{X_{2r+1} = 2k+1\} \leq \frac{4}{\sqrt{\pi}}\frac{1}{\sqrt{2r+1}}. \tag{2.24}$$

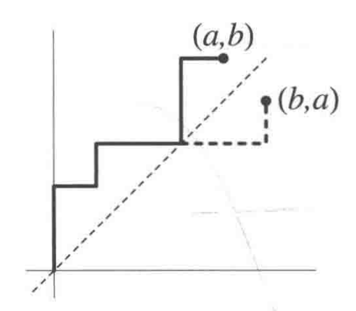


FIGURE 2.4. For the Ballot Theorem: reflecting a "bad" path after the first time the vote counts are equal yields a path to (b, a).

Putting together (2.23) and (2.24) establishes (2.22), since $4/\sqrt{\pi} \leq 3$. ■

PROOF OF THEOREM 2.17. Combining Lemma 2.21 and Lemma 2.22, we obtain (2.19). ■

2.7.1. The Ballot Theorem*. The bijection illustrated in Figure 2.3 has another very nice consequence. Define an ***up-right path*** to be a path through the two-dimensional grid in which every segment heads either up or to the right.

THEOREM 2.24 (Ballot Theorem). *Fix positive integers a and b with $a < b$. An up-right path from $(0, 0)$ to (a, b) chosen uniformly at random has probability $\frac{b-a}{a+b}$ of lying strictly above the line $x = y$ (except for its initial point).*

There is a vivid interpretation of Theorem 2.24. Imagine that $a + b$ votes are being tallied. The up-right path graphs the progress of the pair (votes for candidate A, votes for candidate B) as the votes are counted. Assume we are given that the final totals are a votes for A and b votes for B. Then the probability that the winning candidate was always ahead, from the first vote counted to the last, under the assumption that all possible paths leading to these final totals are equally likely, is exactly $(b - a)/(a + b)$.

PROOF. The total number of up-right paths from $(0, 0)$ to (a, b) is $\binom{a+b}{b}$, since there are $a + b$ steps total, of which exactly b steps go right.

How many paths never touch the line $x = y$ after the first step? Any such path must have its first step up, and there are $\binom{a+b-1}{b-1}$ such paths. How many of those paths touch the line $x = y$?

Given a path whose first step is up and that touches the line $x = y$, reflecting the portion after the first touch of $x = y$ yields a path from $(0, 0)$ whose first step is up and which ends at (b, a). See Figure 2.4. Since every up-right path whose first step is up and which ends at (b, a) must cross $x = y$, we obtain every such path via this reflection. Hence there are $\binom{a+b-1}{b}$ "bad" paths to subtract, and the desired probability is

$$\frac{\binom{a+b-1}{b-1} - \binom{a+b-1}{b}}{\binom{a+b}{b}} = \frac{a!b!}{(a+b)!}\left(\frac{(a+b-1)!}{a!(b-1)!} - \frac{(a+b-1)!}{(a-1)!b!}\right) = \frac{b-a}{a+b}.$$

■

REMARK 2.25. Figures 2.3 and 2.4 clearly illustrate versions of the same bijection. The key step in the proof of Theorem 2.24, counting the "bad" paths, is a case of (2.20): look at the paths after their first step, and set $k = 1$, $r = a + b - 1$ and $j = b - a$.

Exercises

EXERCISE 2.1. Show that the system of equations for $0 < k < n$

$$f_k = \frac{1}{2}(1 + f_{k+1}) + \frac{1}{2}(1 + f_{k-1}), \tag{2.25}$$

together with the boundary conditions $f_0 = f_n = 0$ has a unique solution $f_k = k(n-k)$.

Hint: One approach is to define $\Delta_k = f_k - f_{k-1}$ for $1 \leq k \leq n$. Check that $\Delta_k = \Delta_{k+1} + 2$ (so the Δ_k's form an arithmetic progression) and that $\sum_{k=1}^{n} \Delta_k = 0$.

EXERCISE 2.2. Consider a hesitant gambler: at each time, she flips a coin with probability p of success. If it comes up heads, she places a fair one dollar bet. If tails, she does nothing that round, and her fortune stays the same. If her fortune ever reaches 0 or n, she stops playing. Assuming that her initial fortune is k, find the expected number of rounds she will play, in terms of n, k, and p.

EXERCISE 2.3. Consider a random walk on the path $\{0, 1, \dots, n\}$ in which the walk moves left or right with equal probability except when at n and 0. At n, it remains at n with probability 1/2 and moves to $n - 1$ with probability 1/2, and once the walk hits 0, it remains there forever. Compute the expected time of the walk's absorption at state 0, given that it starts at state n.

EXERCISE 2.4. By comparing the integral of $1/x$ with its Riemann sums, show that

$$\log n \leq \sum_{k=1}^{n} k^{-1} \leq \log n + 1. \tag{2.26}$$

EXERCISE 2.5. Let P be the transition matrix for the Ehrenfest chain described in (2.8). Show that the binomial distribution with parameters n and 1/2 is the stationary distribution for this chain.

EXERCISE 2.6. Give an example of a random walk on a finite abelian group which is *not* reversible.

EXERCISE 2.7. Show that if a random walk on a group is reversible, then the increment distribution is symmetric.

EXERCISE 2.8. Show that when the transition matrix P of a Markov chain is transitive, then the transition matrix $\widehat{P}$ of its time reversal is also transitive.

EXERCISE 2.9. Fix $n \geq 1$. Show that simple random walk on the n-cycle, defined in Example 1.4, is a projection (in the sense of Section 2.3.1) of the simple random walk on $\mathbb{Z}$ defined in Section 2.7.

EXERCISE 2.10 (Reflection Principle). Let (S_n) be the simple random walk on $\mathbb{Z}$. Show that

$$\mathbf{P}\left\{\max_{1 \leq j \leq n} |S_j| \geq c\right\} \leq 2\mathbf{P}\left\{|S_n| \geq c\right\}.$$

Notes

Many of the examples in this chapter are also discussed in Feller (1968). See Chapter XIV for the gambler's ruin, Section IX.3 for coupon collecting, Section V.2 for urn models, and Chapter III for the reflection principle. Grinstead and Snell (1997, Chapter 12) discusses gambler's ruin.

See any undergraduate algebra book, for example Herstein (1975) or Artin (1991), for more information on groups. Much more can be said about random walks on groups than for general Markov chains. Diaconis (1988) is a starting place.

Pólya's urn was introduced in Eggenberger and Pólya (1923) and Pólya (1931). Urns are fundamental models for reinforced processes. See Pemantle (2007) for a wealth of information and many references on urn processes and more generally processes with reinforcement. The book Johnson and Kotz (1977) is devoted to urn models.

See Stanley (1999, pp. 219–229) and Stanley (2008) for many interpretations of the Catalan numbers.

Complements. Generalizations of Theorem 2.17 to walks on $\mathbb{Z}$ other than simple random walks are very useful; we include one here.

THEOREM 2.26. *Let (Δ_i) be i.i.d. integer-valued variables with mean zero and variance σ^2. Let $X_t = \sum_{i=1}^{t} \Delta_i$. Then*

$$\mathbf{P}\{X_t \neq 0 \text{ for } 1 \leq t \leq r\} \leq \frac{4\sigma}{\sqrt{r}}. \tag{2.27}$$

REMARK 2.27. The constant in this estimate is not sharp, but we will give a very elementary proof based on Chebyshev's inequality.

PROOF. For $I \subseteq \mathbb{Z}$, let

$$L_r(I) := \{t \in \{0, 1, \ldots, r\} : X_t \in I\}$$

be the set of times up to and including r when the walk visits I, and write $L_r(v) = L_r(\{v\})$. Define also

$$A_r := \{t \in L_r(0) : X_{t+u} \neq 0 \text{ for } 1 \leq u \leq r\},$$

the set of times t in $L_r(0)$ where the walk does not visit 0 for r steps after t. Since the future of the walk after visiting 0 is independent of the walk up until this time,

$$\mathbf{P}\{t \in A_r\} = \mathbf{P}\{t \in L_r(0)\}\alpha_r,$$

where

$$\alpha_r := \mathbf{P}_0\{X_t \neq 0,\ t = 1, \ldots, r\}.$$

Summing this over $t \in \{0, 1, \ldots, r\}$ and noting that $|A_r| \leq 1$ gives

$$1 \geq \mathbf{E}|A_r| = \mathbf{E}|L_r(0)|\alpha_r. \tag{2.28}$$

It remains to estimate $\mathbf{E}|L_r(0)|$ from below, and this can be done using the local Central Limit Theorem or (in special cases) Stirling's formula.

A more direct (but less precise) approach is to first use Chebyshev's inequality to show that

$$\mathbf{P}\{|X_t| \geq \sigma\sqrt{r}\} \leq \frac{t}{r}$$

and then deduce for $I = (-\sigma\sqrt{r}, \sigma\sqrt{r})$ that

$$\mathbf{E}|L_r(I^c)| \le \sum_{t=1}^{r} \frac{t}{r} = \frac{r+1}{2},$$

whence $\mathbf{E}|L_r(I)| \ge r/2$. For any $v \neq 0$, we have

$$\mathbf{E}|L_r(v)| = \mathbf{E}\left(\sum_{t=0}^{r} \mathbf{1}_{\{X_t = v\}}\right) = \mathbf{E}\left(\sum_{t=\tau_v}^{r} \mathbf{1}_{\{X_t = v\}}\right). \tag{2.29}$$

By the Markov property, the chain after time τ_v has the same distribution as the chain started from v. Hence the right-hand side of (2.29) is bounded above by

$$\mathbf{E}_v\left(\sum_{t=0}^{r} \mathbf{1}_{\{X_t = v\}}\right) = \mathbf{E}_0\left(\sum_{t=0}^{r} \mathbf{1}_{\{X_t = 0\}}\right).$$

We conclude that $r/2 \le \mathbf{E}|L_r(I)| \le 2\sigma\sqrt{r}\mathbf{E}|L_r(0)|$. Thus $\mathbf{E}|L_r(0)| \ge \sqrt{r}/(4\sigma)$. In conjunction with (2.28) this proves (2.27). ■

COROLLARY 2.28. *For the lazy simple random walk on $\mathbb{Z}$ started at height k,*

$$\mathbf{P}_k\{\tau_0^+ > r\} \le \frac{8k}{\sqrt{r}}. \tag{2.30}$$

PROOF. By conditioning on the first move of the walk and then using the fact that the distribution of the walk is symmetric about 0, for $r \ge 1$,

$$\mathbf{P}_0\{\tau_0^+ > r\} = \frac{1}{4}\mathbf{P}_1\{\tau_0^+ > r-1\} + \frac{1}{4}\mathbf{P}_{-1}\{\tau_0^+ > r-1\} = \frac{1}{2}\mathbf{P}_1\{\tau_0^+ > r-1\}. \tag{2.31}$$

Note that when starting from 1, the event that the walk hits height k before visiting 0 for the first time and subsequently does not hit 0 for r steps is contained in the event that the walk started from 1 does not hit 0 for $r-1$ steps. Thus, from (2.31) and Theorem 2.26,

$$\mathbf{P}_1\{\tau_k < \tau_0\}\mathbf{P}_k\{\tau_0^+ > r\} \le \mathbf{P}_1\{\tau_0 > r-1\} = 2\mathbf{P}_0\{\tau_0^+ > r\} \le \frac{8}{\sqrt{r}}. \tag{2.32}$$

(The variance σ^2 of the increments of the lazy random walk is $1/2$, which we bound by 1.) From the gambler's ruin formula given in (2.1), the chance that a simple random walk starting from height 1 hits k before visiting 0 is $1/k$. The probability is the same for a lazy random walk, so together with (2.32) this implies (2.30). ■

CHAPTER 3

Markov Chain Monte Carlo: Metropolis and Glauber Chains

3.1. Introduction

Given an irreducible transition matrix P, there is a unique stationary distribution π satisfying $\pi = \pi P$, which we constructed in Section 1.5. We now consider the inverse problem: given a probability distribution π on Ω, can we find a transition matrix P for which π is its stationary distribution? The following example illustrates why this is a natural problem to consider.

A ***random sample*** from a finite set Ω will mean a random uniform selection from Ω, i.e., one such that each element has the same chance $1/|\Omega|$ of being chosen.

Fix a set $\{1, 2, \ldots, q\}$ of *colors*. A ***proper q-coloring*** of a graph $G = (V, E)$ is an assignment of colors to the vertices V, subject to the constraint that neighboring vertices do not receive the same color. There are (at least) two reasons to look for an efficient method to sample from Ω, the set of all proper q-colorings. If a random sample can be produced, then the size of Ω can be estimated (as we discuss in detail in Section 14.4.2). Also, if it is possible to sample from Ω, then average characteristics of colorings can be studied via simulation.

For some graphs, e.g. trees, there are simple recursive methods for generating a random proper coloring (see Example 14.10). However, for other graphs it can be challenging to directly construct a random sample. One approach is to use Markov chains to sample: suppose that (X_t) is a chain with state space Ω and with stationary distribution uniform on Ω (in Section 3.3, we will construct one such chain). By the Convergence Theorem (Theorem 4.9, whose proof we have not yet given but have often foreshadowed), X_t is approximately uniformly distributed when t is large.

This method of sampling from a given probability distribution is called ***Markov chain Monte Carlo***. Suppose π is a probability distribution on Ω. If a Markov chain (X_t) with stationary distribution π can be constructed, then, for t large enough, the distribution of X_t is close to π. The focus of this book is to determine how large t must be to obtain a sufficient approximation. In this chapter we will focus on the task of finding chains with a given stationary distribution.

3.2. Metropolis Chains

Given *some* chain with state space Ω and an arbitrary stationary distribution, can the chain be modified so that the new chain has the stationary distribution π? The Metropolis algorithm accomplishes this.

3.2.1. Symmetric base chain. Suppose that Ψ is a symmetric transition matrix. In this case, Ψ is reversible with respect to the uniform distribution on Ω.

We now show how to modify transitions made according to Ψ to obtain a chain with stationary distribution π, where π is any probability distribution on Ω.

The new chain evolves as follows: when at state x, a candidate move is generated from the distribution $\Psi(x,\cdot)$. If the proposed new state is y, then the move is censored with probability $1-a(x,y)$. That is, with probability $a(x,y)$, the state y is "accepted" so that the next state of the chain is y, and with the remaining probability $1-a(x,y)$, the chain remains at x. Rejecting moves slows the chain and can reduce its computational efficiency but may be necessary to achieve a specific stationary distribution. We will discuss how to choose the acceptance probability $a(x,y)$ below, but for now observe that the transition matrix P of the new chain is

$$P(x,y) = \begin{cases} \Psi(x,y)a(x,y) & \text{if } y \neq x, \\ 1 - \sum\limits_{z\,:\,z\neq x} \Psi(x,z)a(x,z) & \text{if } y = x. \end{cases}$$

By Proposition 1.19, the transition matrix P has stationary distribution π if

$$\pi(x)\Psi(x,y)a(x,y) = \pi(y)\Psi(y,x)a(y,x) \tag{3.1}$$

for all $x \neq y$. Since we have assumed Ψ is symmetric, equation (3.1) holds if and only if

$$b(x,y) = b(y,x), \tag{3.2}$$

where $b(x,y) = \pi(x)a(x,y)$. Because $a(x,y)$ is a probability and must satisfy $a(x,y) \leq 1$, the function b must obey the constraints

$$\begin{aligned} b(x,y) &\leq \pi(x), \\ b(x,y) = b(y,x) &\leq \pi(y). \end{aligned} \tag{3.3}$$

Since rejecting the moves of the original chain Ψ is wasteful, a solution b to (3.2) and (3.3) should be chosen which is as large as possible. Clearly, all solutions are bounded above by $b(x,y) = \pi(x) \wedge \pi(y) := \min\{\pi(x), \pi(y)\}$. For this choice, the acceptance probability $a(x,y)$ is equal to $(\pi(y)/\pi(x)) \wedge 1$.

The ***Metropolis chain*** for a probability π and a symmetric transition matrix Ψ is defined as

$$P(x,y) = \begin{cases} \Psi(x,y)\left[1 \wedge \frac{\pi(y)}{\pi(x)}\right] & \text{if } y \neq x, \\ 1 - \sum_{z\,:\,z\neq x} \Psi(x,z)\left[1 \wedge \frac{\pi(z)}{\pi(x)}\right] & \text{if } y = x. \end{cases}$$

Our discussion above shows that π is indeed a stationary distribution for the Metropolis chain.

REMARK 3.1. A very important feature of the Metropolis chain is that it only depends on the ratios $\pi(x)/\pi(y)$. Frequently $\pi(x)$ has the form $h(x)/Z$, where the function $h : \Omega \to [0,\infty)$ is known and $Z = \sum_{x\in\Omega} h(x)$ is a normalizing constant. It may be difficult to explicitly compute Z, especially if Ω is large. Because the Metropolis chain only depends on $h(x)/h(y)$, it is not necessary to compute the constant Z in order to simulate the chain. The optimization chains described below (Example 3.2) are examples of this type.

EXAMPLE 3.2 (Optimization). Let f be a real-valued function defined on the vertex set Ω of a graph. In many applications it is desirable to find a vertex x where $f(x)$ is maximal. If the domain Ω is very large, then an exhaustive search may be too expensive.

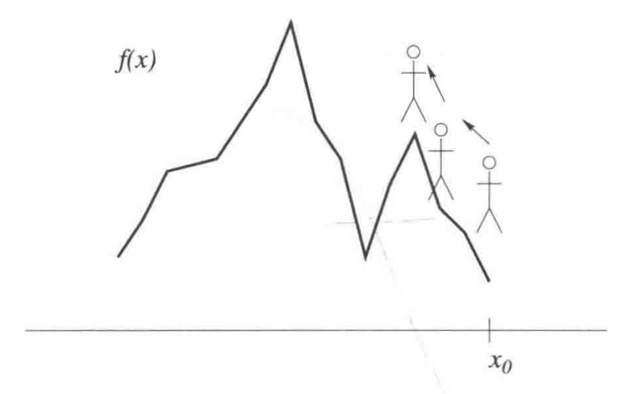


FIGURE 3.1. A hill climb algorithm may become trapped at a local maximum.

A ***hill climb*** is an algorithm which attempts to locate the maximum values of f as follows: when at x, if a neighbor y of x has $f(y) > f(x)$, move to y. When f has local maxima that are not also global maxima, then the climber may become trapped before discovering a global maximum—see Figure 3.1.

One solution is to randomize moves so that instead of always remaining at a local maximum, with some probability the climber moves to lower states.

Suppose for simplicity that Ω is a regular graph, so that simple random walk on Ω has a symmetric transition matrix. Fix $\lambda \geq 1$ and define

$$\pi_\lambda(x) = \frac{\lambda^{f(x)}}{Z(\lambda)},$$

where $Z(\lambda) := \sum_{x\in\Omega} \lambda^{f(x)}$ is the normalizing constant that makes π_λ a probability measure (as mentioned in Remark 3.1, running the Metropolis chain does not require computation of $Z(\lambda)$, which may be prohibitively expensive to compute). Since $\pi_\lambda(x)$ is increasing in $f(x)$, the measure π_λ favors vertices x for which $f(x)$ is large.

If $f(y) < f(x)$, the Metropolis chain accepts a transition $x \to y$ with probability $\lambda^{-[f(x)-f(y)]}$. As $\lambda \to \infty$, the chain more closely resembles the deterministic hill climb.

Define

$$\Omega^\star := \left\{ x \in \Omega \, : \, f(x) = f^\star := \max_{y\in\Omega} f(y) \right\}.$$

Then

$$\lim_{\lambda\to\infty} \pi_\lambda(x) = \lim_{\lambda\to\infty} \frac{\lambda^{f(x)}/\lambda^{f^\star}}{|\Omega^\star| + \sum_{x\in\Omega\setminus\Omega^\star} \lambda^{f(x)}/\lambda^{f^\star}} = \frac{\mathbf{1}_{\{x\in\Omega^\star\}}}{|\Omega^\star|}.$$

That is, as $\lambda \to \infty$, the stationary distribution π_λ of this Metropolis chain converges to the uniform distribution over the global maxima of f.

3.2.2. General base chain. The Metropolis chain can also be defined when the initial transition matrix is not symmetric. For a general (irreducible) transition matrix Ψ and an arbitrary probability distribution π on Ω, the Metropolized chain is executed as follows. When at state x, generate a state y from $\Psi(x, \cdot)$. Move to

y with probability

$$\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1, \tag{3.4}$$

and remain at x with the complementary probability. The transition matrix P for this chain is

$$P(x,y) = \begin{cases} \Psi(x,y)\left[\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1\right] & \text{if } y \neq x, \\ 1 - \sum\limits_{z\,:\,z\neq x} \Psi(x,z)\left[\frac{\pi(z)\Psi(z,x)}{\pi(x)\Psi(x,z)} \wedge 1\right] & \text{if } y = x. \end{cases} \tag{3.5}$$

The reader should check that the transition matrix (3.5) defines a reversible Markov chain with stationary distribution π (see Exercise 3.1).

EXAMPLE 3.3. Suppose you know neither the vertex set V nor the edge set E of a graph G. However, you are able to perform a simple random walk on G. (Many computer and social networks have this form; each vertex knows who its neighbors are, but not the global structure of the graph.) If the graph is not regular, then the stationary distribution is not uniform, so the distribution of the walk will not converge to uniform. You desire a uniform sample from V. We can use the Metropolis algorithm to modify the simple random walk and ensure a uniform stationary distribution. The acceptance probability in (3.4) reduces in this case to

$$\frac{\deg(x)}{\deg(y)} \wedge 1.$$

This biases the walk against moving to higher degree vertices, giving a uniform stationary distribution. Note that it is not necessary to know the size of the vertex set to perform this modification, which can be an important consideration in applications.

3.3. Glauber Dynamics

We will study many chains whose state spaces are contained in a set of the form S^V, where V is the vertex set of a graph and S is a finite set. The elements of S^V, called ***configurations***, are the functions from V to S. We visualize a configuration as a labeling of vertices with elements of S.

Given a probability distribution π on a space of configurations, the Glauber dynamics for π, to be defined below, is a Markov chain which has stationary distribution π. This chain is often called the *Gibbs sampler*, especially in statistical contexts.

3.3.1. Two examples.

As we defined in Section 3.1, a proper q-coloring of a graph $G = (V, E)$ is an element x of $\{1, 2, \ldots, q\}^V$, the set of functions from V to $\{1, 2, \ldots, q\}$, such that $x(v) \neq x(w)$ for all edges $\{v, w\}$. We construct here a Markov chain on the set of proper q-colorings of G.

For a given configuration x and a vertex v, call a color j ***allowable*** at v if j is different from all colors assigned to neighbors of v. That is, a color is allowable at v if it does *not* belong to the set $\{x(w) : w \sim v\}$. Given a proper q-coloring x, we can generate a new coloring by

- selecting a vertex $v \in V$ at random,
- selecting a color j uniformly at random from the allowable colors at v, and

- re-coloring vertex v with color j.

We claim that the resulting chain has uniform stationary distribution: why? Note that transitions are permitted only between colorings differing at a single vertex. If x and y agree everywhere except vertex v, then the chance of moving from x to y equals $|V|^{-1}|A_v(x)|^{-1}$, where $A_v(x)$ is the set of allowable colors at v in x. Since $A_v(x) = A_v(y)$, this probability equals the probability of moving from y to x. Since $P(x, y) = P(y, x)$, the detailed balance equations are satisfied by the uniform distribution.

This chain is called the ***Glauber dynamics for proper q-colorings***. Note that when a vertex v is updated in coloring x, a coloring is chosen from π conditioned on the set of colorings agreeing with x at all vertices different from v. This is the general rule for defining Glauber dynamics for any set of configurations. Before spelling out the details in the general case, we consider one other specific example.

A ***hardcore configuration*** is a placement of particles on the vertices V of a graph so that each vertex is occupied by at most one particle and no two particles are adjacent. Formally, a hardcore configuration x is an element of $\{0, 1\}^V$, the set of functions from V to $\{0, 1\}$, satisfying $x(v)x(w) = 0$ whenever v and w are neighbors. The vertices v with $x(v) = 1$ are called ***occupied***, and the vertices v with $x(v) = 0$ are called ***vacant***.

Consider the following transition rule:

- a vertex v is chosen uniformly at random, and, regardless of the current status of v,
- if any neighbor of v is occupied, v is left unoccupied, while if no adjacent vertex is occupied, a particle is placed at v with probability $1/2$.

REMARK 3.4. Note that the rule above has the same effect as the following apparently simpler rule: if no neighbor of v is occupied, then, with probability $1/2$, flip the status of v. Our original description will be much more convenient when, in the future, we attempt to couple multiple copies of this chain, since it provides a way to ensure that the status at the chosen vertex v is the same in all copies after an update. See Section 5.4.2.

The verification that this chain is reversible with respect to the uniform distribution is similar to the coloring chain just considered and is left to the reader.

3.3.2. General definition.

In general, let V and S be finite sets, and suppose that Ω is a subset of S^V (both the set of proper q-colorings and the set of hardcore configurations are of this form). Let π be a probability distribution whose support is Ω. The (single-site) ***Glauber dynamics for*** π is a reversible Markov chain with state space Ω, stationary distribution π, and the transition probabilities we describe below.

In words, the Glauber chain moves from state x as follows: a vertex v is chosen uniformly at random from V, and a new state is chosen according to the measure π conditioned on the set of states equal to x at all vertices different from v. We give the details now. For $x \in \Omega$ and $v \in V$, let

$$\Omega(x, v) = \{y \in \Omega : y(w) = x(w) \text{ for all } w \neq v\} \tag{3.6}$$

be the set of states agreeing with x everywhere except possibly at v, and define

$$\pi^{x,v}(y) = \pi(y \mid \Omega(x,v)) = \begin{cases} \frac{\pi(y)}{\pi(\Omega(x,v))} & \text{if } y \in \Omega(x,v), \\ 0 & \text{if } y \notin \Omega(x,v) \end{cases}$$

to be the distribution π conditioned on the set $\Omega(x,v)$. The rule for updating a configuration x is: pick a vertex v uniformly at random, and choose a new configuration according to $\pi^{x,v}$.

The distribution π is always stationary and reversible for the Glauber dynamics (see Exercise 3.2).

3.3.3. Comparing Glauber dynamics and Metropolis chains. Suppose now that π is a probability distribution on the state space S^V, where S is a finite set and V is the vertex set of a graph. We can always define the Glauber chain as just described. Suppose on the other hand that we have a chain which picks a vertex v at random and has *some* mechanism for updating the configuration at v. (For example, the chain may pick an element of S at random to update at v.) This chain may not have stationary distribution π, but it can be modified by the Metropolis rule to obtain a chain with stationary distribution π. This chain can be very similar to the Glauber chain, but may not coincide exactly. We consider our examples.

EXAMPLE 3.5 (Chains on q-colorings). Consider the following chain on (not necessarily proper) q-colorings: a vertex v is chosen uniformly at random, a color is selected uniformly at random among *all* q colors, and the vertex v is recolored with the chosen color. We apply the Metropolis rule to this chain, where π is the probability measure which is uniform over the space of *proper* q-colorings. When at a proper coloring, if the color k is proposed to update a vertex, then the Metropolis rule accepts the proposed re-coloring with probability 1 if it yields a proper coloring and rejects otherwise.

The Glauber chain described in Section 3.3.1 is slightly different. Note in particular that the chance of remaining at the same coloring differs for the two chains. If there are a allowable colors at vertex v and this vertex v is selected for updating in the Glauber dynamics, the chance that the coloring remains the same is $1/a$. For the Metropolis chain, if vertex v is selected, the chance of remaining in the current coloring is $(1+q-a)/q$.

EXAMPLE 3.6 (Hardcore chains). Again identify elements of $\{0,1\}^V$ with a placement of particles onto the vertex set V, and consider the following chain on $\{0,1\}^V$: a vertex is chosen at random, and a particle is placed at the selected vertex with probability $1/2$. This chain does not live on the space of hardcore configurations, as there is no constraint against placing a particle on a vertex with an occupied neighbor.

We can modify this chain with the Metropolis rule to obtain a chain with stationary distribution π, where π is uniform over hardcore configurations. If x is a hardcore configuration, the move $x \to y$ is rejected if and only if y is not a hardcore configuration. The Metropolis chain and the Glauber dynamics agree in this example.

3.3.4. Hardcore model with fugacity. Let $G = (V, E)$ be a graph and let Ω be the set of hardcore configurations on G. The ***hardcore model*** with ***fugacity***

λ is the probability π on hardcore configurations defined by

$$\pi(\sigma) = \begin{cases} \frac{\lambda^{\sum_{v \in V} \sigma(v)}}{Z(\lambda)} & \text{if } \sigma(v)\sigma(w) = 0 \text{ for all } \{v, w\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The factor $Z(\lambda) = \sum_{\sigma \in \Omega} \lambda^{\sum_{v \in V} \sigma(v)}$ normalizes π to have unit total mass.

The Glauber dynamics for the hardcore model updates a configuration $X_t = \sigma$ to a new configuration X_{t+1} as follows: a vertex w is chosen at random. Denote the set of occupied neighbors of w by $\mathcal{N}$, so that

$$\mathcal{N}(w) := \{v : v \sim w \text{ and } \sigma(v) = 1\}.$$

If $\mathcal{N}(w) \neq \varnothing$, then $X_{t+1} = \sigma$. If $\mathcal{N}(w) = \varnothing$, then set

$$X_{t+1}(w) = \begin{cases} 1 & \text{with probability } \lambda/(1+\lambda), \\ 0 & \text{with probability } 1/(1+\lambda). \end{cases}$$

Set $X_{t+1}(v) = \sigma(v)$ for all $v \neq w$.

3.3.5. The Ising model. A ***spin system*** is a probability distribution on $\Omega = \{-1, 1\}^V$, where V is the vertex set of a graph $G = (V, E)$. The value $\sigma(v)$ is called the ***spin*** at v. The physical interpretation is that magnets, each having one of the two possible orientations represented by $+1$ and -1, are placed on the vertices of the graph; a configuration specifies the orientations of these magnets.

The nearest-neighbor ***Ising model*** is the most widely studied spin system. In this system, the ***energy*** of a configuration σ is defined to be

$$H(\sigma) = -\sum_{\substack{v,w \in V \\ v \sim w}} \sigma(v)\sigma(w). \tag{3.7}$$

Clearly, the energy increases with the number of pairs of neighbors whose spins disagree (anyone who has played with magnets has observed firsthand that it is challenging to place neighboring magnets in opposite orientations and keep them there).

The ***Gibbs distribution*** corresponding to the energy H is the probability distribution μ on Ω defined by

$$\mu(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}. \tag{3.8}$$

Here the ***partition function*** $Z(\beta)$ is the normalizing constant required to make μ a probability distribution:

$$Z(\beta) := \sum_{\sigma \in \Omega} e^{-\beta H(\sigma)}. \tag{3.9}$$

The parameter $\beta \geq 0$ determines the importance of the energy function. In the physical interpretation, β is the reciprocal of temperature. At infinite temperature ($\beta = 0$), the energy function H plays no role and μ is the uniform distribution on Ω. In this case, there is no interaction between the spins at differing vertices and the random variables $\{\sigma(v)\}_{v \in V}$ are independent. As $\beta > 0$ increases, the bias of μ towards low-energy configurations also increases. See Figure 3.2 for an illustration of the effect of β on configurations.

The Glauber dynamics for the Gibbs distribution μ move from a starting configuration σ by picking a vertex w uniformly at random from V and then generating

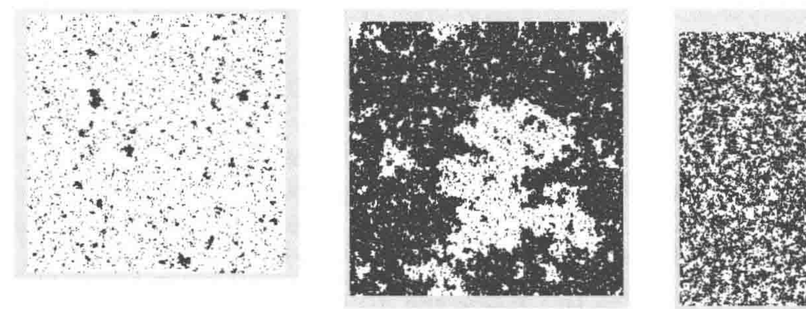
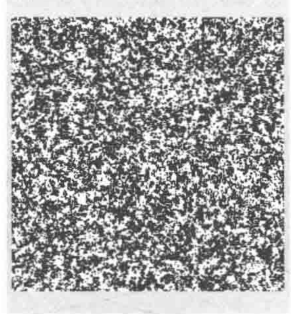



FIGURE 3.2. Glauber dynamics for the Ising model on the 250 × 250 torus viewed at times $t = 1{,}000$, 16,500, and 1,000 at low, critical, and high temperature, respectively. Simulations and graphics courtesy of Raissa D'Souza.

a new configuration according to μ conditioned on the set of configurations agreeing with σ on vertices different from w.

The reader can check that the conditional μ-probability of spin +1 at w is

$$p(\sigma, w) := \frac{e^{\beta S(\sigma,w)}}{e^{\beta S(\sigma,w)} + e^{-\beta S(\sigma,w)}} = \frac{1 + \tanh(\beta S(\sigma, w))}{2}, \tag{3.10}$$

where $S(\sigma, w) := \sum_{u\,:\,u\sim w} \sigma(u)$. Note that $p(\sigma, w)$ depends only on the spins at vertices adjacent to w. Therefore, the transition matrix on Ω is given by

$$P(\sigma, \sigma') = \frac{1}{|V|} \sum_{v \in V} \frac{e^{\beta\, \sigma'(w)\, S(\sigma,w)}}{e^{\beta\, \sigma'(w)\, S(\sigma,w)} + e^{-\beta\, \sigma'(w)\, S(\sigma,w)}} \cdot \mathbf{1}_{\{\sigma(v)=\sigma'(v) \text{ for } v \neq w\}}. \tag{3.11}$$

This chain has stationary distribution given by the Gibbs distribution μ.

Exercises

EXERCISE 3.1. Let Ψ be an irreducible transition matrix on Ω, and let π be a probability distribution on Ω. Show that the transition matrix

$$P(x, y) = \begin{cases} \Psi(x, y) \left[\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1\right] & \text{if } y \neq x, \\ 1 - \sum\limits_{z\,:\,z \neq x} \Psi(x, z) \left[\frac{\pi(z)\Psi(z,x)}{\pi(x)\Psi(x,z)} \wedge 1\right] & \text{if } y = x \end{cases}$$

defines a reversible Markov chain with stationary distribution π.

EXERCISE 3.2. Verify that the Glauber dynamics for π is a reversible Markov chain with stationary distribution π.

Notes

The Metropolis chain was introduced in Metropolis, Rosenbluth, Teller, and Teller (1953) for a specific stationary distribution. Hastings (1970) extended the

method to general chains and distributions. The survey by Diaconis and Saloff-Coste (1998) contains more on the Metropolis algorithm. The textbook by Brémaud (1999) also discusses the use of the Metropolis algorithm for Monte Carlo sampling.

Variations on the randomized hill climb in Example 3.2 used to locate extrema, especially when the parameter λ is tuned as the walk progresses, are called *simulated annealing* algorithms. Significant references are Holley and Stroock (1988) and Hajek (1988).

We will have much more to say about Glauber dynamics for colorings in Section 14.3 and about Glauber dynamics for the Ising model in Chapter 15.

Häggström (2007) proves interesting inequalities using the Markov chains of this chapter.

CHAPTER 4

Introduction to Markov Chain Mixing

We are now ready to discuss the long-term behavior of finite Markov chains. Since we are interested in quantifying the speed of convergence of families of Markov chains, we need to choose an appropriate metric for measuring the distance between distributions.

First we define ***total variation distance*** and give several characterizations of it, all of which will be useful in our future work. Next we prove the Convergence Theorem (Theorem 4.9), which says that for an irreducible and aperiodic chain the distribution after many steps approaches the chain's stationary distribution, in the sense that the total variation distance between them approaches 0. In the rest of the chapter we examine the effects of the initial distribution on distance from stationarity, define the *mixing time* of a chain, consider circumstances under which related chains can have identical mixing, and prove a version of the Ergodic Theorem (Theorem 4.16) for Markov chains.

4.1. Total Variation Distance

The ***total variation distance*** between two probability distributions μ and ν on Ω is defined by

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|. \tag{4.1}$$

This definition is explicitly probabilistic: the distance between μ and ν is the maximum difference between the probabilities assigned to a single event by the two distributions.

EXAMPLE 4.1. Recall the coin-tossing frog of Example 1.1, who has probability p of jumping from east to west and probability q of jumping from west to east. His transition matrix is $\begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ and his stationary distribution is $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$. Assume the frog starts at the east pad (that is, $\mu_0 = (1, 0)$) and define

$$\Delta_t = \mu_t(e) - \pi(e).$$

Since there are only two states, there are only four possible events $A \subseteq \Omega$. Hence it is easy to check (and you should) that

$$\|\mu_t - \pi\|_{\mathrm{TV}} = \Delta_t = P^t(e,e) - \pi(e) = \pi(w) - P^t(e,w).$$

We pointed out in Example 1.1 that $\Delta_t = (1-p-q)^t \Delta_0$. Hence for this two-state chain, the total variation distance decreases exponentially fast as t increases. (Note that $(1-p-q)$ is an eigenvalue of P; we will discuss connections between eigenvalues and mixing in Chapter 12.)

The definition of total variation distance (4.1) is a maximum over *all* subsets of Ω, so using this definition is not always the most convenient way to estimate

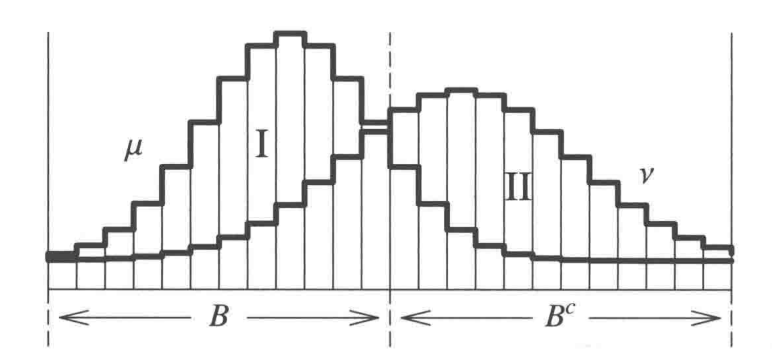


FIGURE 4.1. Recall that $B = \{x : \mu(x) > \nu(x)\}$. Region I has area $\mu(B) - \nu(B)$. Region II has area $\nu(B^c) - \mu(B^c)$. Since the total area under each of μ and ν is 1, regions I and II must have the same area—and that area is $\|\mu - \nu\|_{TV}$.

the distance. We now give three extremely useful alternative characterizations. Proposition 4.2 reduces total variation distance to a simple sum over the state space. Proposition 4.7 uses *coupling* to give another probabilistic interpretation: $\|\mu - \nu\|_{TV}$ measures how close to identical we can force two random variables realizing μ and ν to be.

PROPOSITION 4.2. *Let μ and ν be two probability distributions on Ω. Then*

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| . \tag{4.2}$$

PROOF. Let $B = \{x : \mu(x) \geq \nu(x)\}$ and let $A \subset \Omega$ be any event. Then

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B). \tag{4.3}$$

The first inequality is true because any $x \in A \cap B^c$ satisfies $\mu(x) - \nu(x) < 0$, so the difference in probability cannot decrease when such elements are eliminated. For the second inequality, note that including more elements of B cannot decrease the difference in probability.

By exactly parallel reasoning,

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c). \tag{4.4}$$

Fortunately, the upper bounds on the right-hand sides of (4.3) and (4.4) are actually the same (as can be seen by subtracting them; see Figure 4.1). Furthermore, when we take $A = B$ (or B^c), then $|\mu(A) - \nu(A)|$ is equal to the upper bound. Thus

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \left[\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)\right] = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

■

REMARK 4.3. The proof of Proposition 4.2 also shows that

$$\|\mu - \nu\|_{TV} = \sum_{\substack{x \in \Omega \\ \mu(x) \geq \nu(x)}} [\mu(x) - \nu(x)], \tag{4.5}$$

which is a useful identity.

REMARK 4.4. From Proposition 4.2 and the triangle inequality for real numbers, it is easy to see that total variation distance satisfies the triangle inequality: for probability distributions μ, ν and η,

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \|\mu - \eta\|_{\mathrm{TV}} + \|\eta - \nu\|_{\mathrm{TV}} . \tag{4.6}$$

PROPOSITION 4.5. *Let μ and ν be two probability distributions on Ω. Then the total variation distance between them satisfies*

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sup \left\{ \sum_{x \in \Omega} f(x)\mu(x) - \sum_{x \in \Omega} f(x)\nu(x) \ : \ f \text{ satisfying } \max_{x \in \Omega} |f(x)| \leq 1 \right\}. \tag{4.7}$$

PROOF. When f satisfies $\max_{x \in \Omega} |f(x)| \leq 1$, we have

$$\begin{aligned} \frac{1}{2} \left| \sum_{x \in \Omega} f(x)\mu(x) - \sum_{x \in \Omega} f(x)\nu(x) \right| &\leq \frac{1}{2} \sum_{x \in \Omega} |f(x)[\mu(x) - \nu(x)]| \\ &\leq \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| \\ &= \|\mu - \nu\|_{\mathrm{TV}} , \end{aligned}$$

which shows that the right-hand side of (4.7) is not more than $\|\mu - \nu\|_{\mathrm{TV}}$. Define

$$f^{\star}(x) = \begin{cases} 1 & \text{if } x \text{ satisfies } \mu(x) \geq \nu(x), \\ -1 & \text{if } x \text{ satisfies } \mu(x) < \nu(x). \end{cases}$$

Then

$$\begin{aligned} \frac{1}{2} \left[\sum_{x \in \Omega} f^{\star}(x)\mu(x) - \sum_{x \in \Omega} f^{\star}(x)\nu(x) \right] &= \frac{1}{2} \sum_{x \in \Omega} f^{\star}(x)[\mu(x) - \nu(x)] \\ &= \frac{1}{2} \left[\sum_{\substack{x \in \Omega \\ \mu(x) \geq \nu(x)}} [\mu(x) - \nu(x)] + \sum_{\substack{x \in \Omega \\ \nu(x) > \mu(x)}} [\nu(x) - \mu(x)] \right]. \end{aligned}$$

Using (4.5) shows that the right-hand side above equals $\|\mu - \nu\|_{\mathrm{TV}}$. Hence the right-hand side of (4.7) is at least $\|\mu - \nu\|_{\mathrm{TV}}$. ■

4.2. Coupling and Total Variation Distance

A ***coupling*** of two probability distributions μ and ν is a pair of random variables (X, Y) defined on a single probability space such that the marginal distribution of X is μ and the marginal distribution of Y is ν. That is, a coupling (X, Y) satisfies $\mathbf{P}\{X = x\} = \mu(x)$ and $\mathbf{P}\{Y = y\} = \nu(y)$.

Coupling is a general and powerful technique; it can be applied in many different ways. Indeed, Chapters 5 and 14 use couplings of entire chain trajectories to bound rates of convergence to stationarity. Here, we offer a gentle introduction by showing the close connection between couplings of two random variables and the total variation distance between those variables.

EXAMPLE 4.6. Let μ and ν both be the "fair coin" measure giving weight $1/2$ to the elements of $\{0, 1\}$.

(i) One way to couple μ and ν is to define (X, Y) to be a pair of independent coins, so that $\mathbf{P}\{X = x, Y = y\} = 1/4$ for all $x, y \in \{0, 1\}$.
(ii) Another way to couple μ and ν is to let X be a fair coin toss and define $Y = X$. In this case, $\mathbf{P}\{X = Y = 0\} = 1/2$, $\mathbf{P}\{X = Y = 1\} = 1/2$, and $\mathbf{P}\{X \neq Y\} = 0$.

Given a coupling (X, Y) of μ and ν, if q is the joint distribution of (X, Y) on $\Omega \times \Omega$, meaning that $q(x, y) = \mathbf{P}\{X = x, Y = y\}$, then q satisfies

$$\sum_{y \in \Omega} q(x, y) = \sum_{y \in \Omega} \mathbf{P}\{X = x, Y = y\} = \mathbf{P}\{X = x\} = \mu(x)$$

and

$$\sum_{x \in \Omega} q(x, y) = \sum_{x \in \Omega} \mathbf{P}\{X = x, Y = y\} = \mathbf{P}\{Y = y\} = \nu(y).$$

Conversely, given a probability distribution q on the product space $\Omega \times \Omega$ which satisfies

$$\sum_{y \in \Omega} q(x, y) = \mu(x) \quad \text{and} \quad \sum_{x \in \Omega} q(x, y) = \nu(y),$$

there is a pair of random variables (X, Y) having q as their joint distribution – and consequently this pair (X, Y) is a coupling of μ and ν. In summary, a coupling can be specified either by a pair of random variables (X, Y) defined on a common probability space or by a distribution q on $\Omega \times \Omega$.

Returning to Example 4.6, the coupling in part (i) could equivalently be specified by the probability distribution q_1 on $\{0, 1\}^2$ given by

$$q_1(x, y) = \frac{1}{4} \quad \text{for all } (x, y) \in \{0, 1\}^2.$$

Likewise, the coupling in part (ii) can be identified with the probability distribution q_2 given by

$$q_2(x, y) = \begin{cases} \frac{1}{2} & \text{if } (x, y) = (0, 0),\ (x, y) = (1, 1), \\ 0 & \text{if } (x, y) = (0, 1),\ (x, y) = (1, 0). \end{cases}$$

Any two distributions μ and ν have an independent coupling. However, when μ and ν are not identical, it will not be possible for X and Y to always have the same value. How close can a coupling get to having X and Y identical? Total variation distance gives the answer.

PROPOSITION 4.7. *Let μ and ν be two probability distributions on Ω. Then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf\left\{\mathbf{P}\{X \neq Y\} : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\right\}. \tag{4.8}$$

REMARK 4.8. We will in fact show that there is a coupling (X, Y) which attains the infimum in (4.8). We will call such a coupling ***optimal***.

PROOF. First, we note that for any coupling (X, Y) of μ and ν and any event $A \subset \Omega$,

$$\mu(A) - \nu(A) = \mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\} \tag{4.9}$$

$$\leq \mathbf{P}\{X \in A, Y \notin A\} \tag{4.10}$$

$$\leq \mathbf{P}\{X \neq Y\}. \tag{4.11}$$

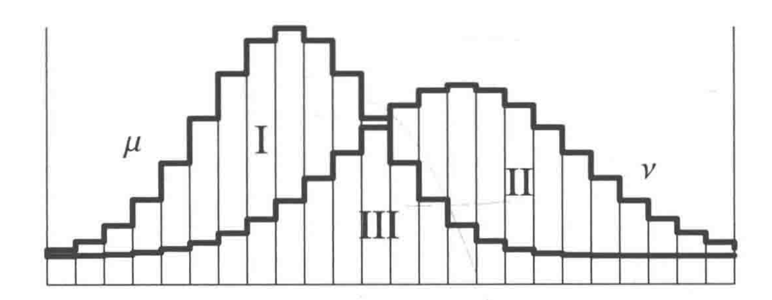


FIGURE 4.2. Since each of regions I and II has area $\|\mu - \nu\|_{\mathrm{TV}}$ and μ and ν are probability measures, region III has area $1 - \|\mu - \nu\|_{\mathrm{TV}}$.

(Dropping the event $\{X \notin A, Y \in A\}$ from the second term of the difference gives the first inequality.) It immediately follows that

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \inf\{\mathbf{P}\{X \neq Y\} : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{4.12}$$

It will suffice to construct a coupling for which $\mathbf{P}\{X \neq Y\}$ is exactly equal to $\|\mu - \nu\|_{\mathrm{TV}}$. We will do so by forcing X and Y to be equal as often as they possibly can be. Consider Figure 4.2. Region III, bounded by $\mu(x) \wedge \nu(x) = \min\{\mu(x), \nu(x)\}$, can be seen as the overlap between the two distributions. Informally, our coupling proceeds by choosing a point in the union of regions I, II, and III. Whenever we "land" in region III, we take $X = Y$. Otherwise, we accept that X must be in region I and Y must be in region II; since those regions have disjoint support, X and Y cannot be equal.

More formally, we use the following procedure to generate X and Y. Let

$$p = \sum_{x \in \Omega} \mu(x) \wedge \nu(x).$$

Write

$$\sum_{x \in \Omega} \mu(x) \wedge \nu(x) = \sum_{\substack{x \in \Omega, \\ \mu(x) \leq \nu(x)}} \mu(x) + \sum_{\substack{x \in \Omega, \\ \mu(x) > \nu(x)}} \nu(x).$$

Adding and subtracting $\sum_{x : \mu(x) > \nu(x)} \mu(x)$ to the right-hand side above shows that

$$\sum_{x \in \Omega} \mu(x) \wedge \nu(x) = 1 - \sum_{\substack{x \in \Omega, \\ \mu(x) > \nu(x)}} [\mu(x) - \nu(x)].$$

By equation (4.5) and the immediately preceding equation,

$$\sum_{x \in \Omega} \mu(x) \wedge \nu(x) = 1 - \|\mu - \nu\|_{\mathrm{TV}} = p. \tag{4.13}$$

Flip a coin with probability of heads equal to p.

(i) If the coin comes up heads, then choose a value Z according to the probability distribution

$$\gamma_{\mathrm{III}}(x) = \frac{\mu(x) \wedge \nu(x)}{p},$$

and set $X = Y = Z$.

(ii) If the coin comes up tails, choose X according to the probability distribution

$$\gamma_{\mathrm{I}}(x) = \begin{cases} \frac{\mu(x)-\nu(x)}{\|\mu-\nu\|_{\mathrm{TV}}} & \text{if } \mu(x) > \nu(x), \\ 0 & \text{otherwise,} \end{cases}$$

and independently choose Y according to the probability distribution

$$\gamma_{\mathrm{II}}(x) = \begin{cases} \frac{\nu(x)-\mu(x)}{\|\mu-\nu\|_{\mathrm{TV}}} & \text{if } \nu(x) > \mu(x), \\ 0 & \text{otherwise.} \end{cases}$$

Note that (4.5) ensures that γ_{I} and γ_{II} are probability distributions.

Clearly,

$$p\gamma_{\mathrm{III}} + (1-p)\gamma_{\mathrm{I}} = \mu,$$
$$p\gamma_{\mathrm{III}} + (1-p)\gamma_{\mathrm{II}} = \nu,$$

so that the distribution of X is μ and the distribution of Y is ν. Note that in the case that the coin lands tails up, $X \neq Y$ since γ_{I} and γ_{II} are positive on disjoint subsets of Ω. Thus $X = Y$ if and only if the coin toss is heads. We conclude that

$$\mathbf{P}\{X \neq Y\} = \|\mu - \nu\|_{\mathrm{TV}}.$$

■

4.3. The Convergence Theorem

We are now ready to prove that irreducible, aperiodic Markov chains converge to their stationary distributions—a key step, as much of the rest of the book will be devoted to estimating the rate at which this convergence occurs. The assumption of aperiodicity is indeed necessary—recall the even n-cycle of Example 1.4.

As is often true of such fundamental facts, there are many proofs of the Convergence Theorem. The one given here decomposes the chain into a mixture of repeated independent sampling from the stationary distribution and another Markov chain. See Exercise 5.1 for another proof using two coupled copies of the chain.

THEOREM 4.9 (Convergence Theorem). *Suppose that P is irreducible and aperiodic, with stationary distribution π. Then there exist constants $\alpha \in (0,1)$ and $C > 0$ such that*

$$\max_{x\in\Omega} \left\| P^t(x,\cdot) - \pi \right\|_{\mathrm{TV}} \leq C\alpha^t. \tag{4.14}$$

PROOF. Since P is irreducible and aperiodic, by Proposition 1.7 there exists an r such that P^r has strictly positive entries. Let Π be the matrix with $|\Omega|$ rows, each of which is the row vector π. For sufficiently small $\delta > 0$, we have

$$P^r(x,y) \geq \delta\pi(y)$$

for all $x, y \in \Omega$. Let $\theta = 1 - \delta$. The equation

$$P^r = (1-\theta)\Pi + \theta Q \tag{4.15}$$

defines a stochastic matrix Q.

It is a straightforward computation to check that $\boldsymbol{M}\Pi = \Pi$ for any stochastic matrix $\boldsymbol{M}$ and that $\Pi\boldsymbol{M} = \Pi$ for any matrix $\boldsymbol{M}$ such that $\pi\boldsymbol{M} = \pi$.

Next, we use induction to demonstrate that

$$P^{rk} = \left(1 - \theta^k\right)\Pi + \theta^k Q^k \tag{4.16}$$

for $k \geq 1$. If $k = 1$, this holds by (4.15). Assuming that (4.16) holds for $k = n$,

$$P^{r(n+1)} = P^{rn}P^r = [(1-\theta^n)\Pi + \theta^n Q^n] P^r. \tag{4.17}$$

Distributing and expanding P^r in the second term (using (4.15)) gives

$$P^{r(n+1)} = [1-\theta^n]\Pi P^r + (1-\theta)\theta^n Q^n \Pi + \theta^{n+1} Q^n Q. \tag{4.18}$$

Using that $\Pi P^r = \Pi$ and $Q^n \Pi = \Pi$ shows that

$$P^{r(n+1)} = \left[1-\theta^{n+1}\right]\Pi + \theta^{n+1}Q^{n+1}. \tag{4.19}$$

This establishes (4.16) for $k = n+1$ (assuming it holds for $k = n$), and hence it holds for all k.

Multiplying by P^j and rearranging terms now yields

$$P^{rk+j} - \Pi = \theta^k \left(Q^k P^j - \Pi\right). \tag{4.20}$$

To complete the proof, sum the absolute values of the elements in row x_0 on both sides of (4.20) and divide by 2. On the right, the second factor is at most the largest possible total variation distance between distributions, which is 1. Hence for any x_0 we have

$$\left\|P^{rk+j}(x_0,\cdot) - \pi\right\|_{TV} \leq \theta^k. \tag{4.21}$$

■

REMARK 4.10. Because of Theorem 4.9, the distribution π is also called the ***equilibrium distribution***.

4.4. Standardizing Distance from Stationarity

Bounding the maximal distance (over $x_0 \in \Omega$) between $P^t(x_0,\cdot)$ and π is among our primary objectives. It is therefore convenient to define

$$d(t) := \max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{TV}. \tag{4.22}$$

We will see in Chapter 5 that it is often possible to bound $\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV}$, uniformly over all pairs of states (x,y). We therefore make the definition

$$\bar{d}(t) := \max_{x,y\in\Omega} \left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV}. \tag{4.23}$$

The relationship between d and $\bar{d}$ is given below:

LEMMA 4.11. *If $d(t)$ and $\bar{d}(t)$ are as defined in* (4.22) *and* (4.23), *respectively, then*

$$d(t) \leq \bar{d}(t) \leq 2d(t). \tag{4.24}$$

PROOF. It is immediate from the triangle inequality for the total variation distance that $\bar{d}(t) \leq 2d(t)$.

To show that $d(t) \leq \bar{d}(t)$, note first that since π is stationary, we have $\pi(A) = \sum_{y\in\Omega}\pi(y)P^t(y,A)$ for any set A. (This is the definition of stationarity if A is a singleton $\{x\}$. To get this for arbitrary A, just sum over the elements in A.) Using this shows that

$$\begin{aligned}
\left\|P^t(x,\cdot) - \pi\right\|_{TV} &= \max_{A\subset\Omega} |P^t(x,A) - \pi(A)| \\
&= \max_{A\subset\Omega} \left| \sum_{y\in\Omega} \pi(y)\left[P^t(x,A) - P^t(y,A)\right] \right|.
\end{aligned}$$

We can use the triangle inequality and the fact that the maximum of a sum is not larger than the sum over a maximum to bound the right-hand side above by

$$\max_{A\subset\Omega}\sum_{y\in\Omega}\pi(y)|P^t(x,A)-P^t(y,A)| \le \sum_{y\in\Omega}\pi(y)\max_{A\subset\Omega}|P^t(x,A)-P^t(y,A)|$$
$$= \sum_{y\in\Omega}\pi(y)\left\|P^t(x,\cdot)-P^t(y,\cdot)\right\|_{TV}. \tag{4.25}$$

Since a weighted average of a set of numbers is never larger than its maximum element, the right-hand side of (4.25) is bounded by $\max_{y\in\Omega}\|P^t(x,\cdot)-P^t(y,\cdot)\|_{TV}$. ■

Let $\mathcal{P}$ denote the collection of all probability distributions on Ω. Exercise 4.1 asks the reader to prove the following equalities:

$$d(t) = \sup_{\mu\in\mathcal{P}}\left\|\mu P^t-\pi\right\|_{TV},$$
$$\bar{d}(t) = \sup_{\mu,\nu\in\mathcal{P}}\left\|\mu P^t-\nu P^t\right\|_{TV}.$$

LEMMA 4.12. *The function $\bar{d}$ is submultiplicative: $\bar{d}(s+t)\le\bar{d}(s)\bar{d}(t)$.*

PROOF. Fix $x,y\in\Omega$, and let (X_s,Y_s) be the optimal coupling of $P^s(x,\cdot)$ and $P^s(y,\cdot)$ whose existence is guaranteed by Proposition 4.7. Hence

$$\|P^s(x,\cdot)-P^s(y,\cdot)\|_{TV}=\mathbf{P}\{X_s\ne Y_s\}.$$

As P^{s+t} is the matrix product of P^t and P^s and the distribution of X_s is $P^s(x,\cdot)$, we have

$$P^{s+t}(x,w)=\sum_z P^s(x,z)P^t(z,w)=\sum_z\mathbf{P}\{X_s=z\}P^t(z,w)=\mathbf{E}\left(P^t(X_s,w)\right). \tag{4.26}$$

Combining this with the similar identity $P^{s+t}(y,w)=\mathbf{E}\left(P^t(Y_s,w)\right)$ allows us to write

$$\begin{aligned}P^{s+t}(x,w)-P^{s+t}(y,w)&=\mathbf{E}\left(P^t(X_s,w)\right)-\mathbf{E}\left(P^t(Y_s,w)\right)\\&=\mathbf{E}\left(P^t(X_s,w)-P^t(Y_s,w)\right).\end{aligned} \tag{4.27}$$

Combining the expectations is possible since X_s and Y_s are defined together on the same probability space.

Summing (4.27) over $w\in\Omega$ and applying Proposition 4.2 shows that

$$\left\|P^{s+t}(x,\cdot)-P^{s+t}(y,\cdot)\right\|_{TV}=\frac{1}{2}\sum_w\left|\mathbf{E}\left(P^t(X_s,w)-P^t(Y_s,w)\right)\right|. \tag{4.28}$$

The right-hand side above is less than or equal to

$$\mathbf{E}\left(\frac{1}{2}\sum_w\left|P^t(X_s,w)-P^t(Y_s,w)\right|\right). \tag{4.29}$$

Applying Proposition 4.2 again, we see that the quantity inside the expectation is exactly the distance $\|P^t(X_s,\cdot)-P^t(Y_s,\cdot)\|_{TV}$, which is zero whenever $X_s=Y_s$. Moreover, this distance is always bounded by $\bar{d}(t)$. This shows that

$$\left\|P^{s+t}(x,\cdot)-P^{s+t}(y,\cdot)\right\|_{TV}\le\bar{d}(t)\mathbf{E}\left(\mathbf{1}_{\{X_s\ne Y_s\}}\right)=\bar{d}(t)\mathbf{P}\{X_s\ne Y_s\}. \tag{4.30}$$

Finally, since (X_s,Y_s) is an optimal coupling, the probability on the right-hand side is equal to $\|P^s(x,\cdot)-P^s(y,\cdot)\|_{TV}$. Maximizing over x,y completes the proof. ■

Exercise 4.3 implies that $\bar{d}(t)$ is non-increasing in t. From this and Lemma 4.12 it follows that when c is any non-negative integer and t is any non-negative integer, we have

$$d(ct) \leq \bar{d}(ct) \leq \bar{d}(t)^c. \tag{4.31}$$

4.5. Mixing Time

It is useful to introduce a parameter which measures the time required by a Markov chain for the distance to stationarity to be small. The ***mixing time*** is defined by

$$t_{\text{mix}}(\varepsilon) := \min\{t \, : \, d(t) \leq \varepsilon\} \tag{4.32}$$

and

$$t_{\text{mix}} := t_{\text{mix}}(1/4). \tag{4.33}$$

Lemma 4.11 and (4.31) show that when ℓ is a non-negative integer,

$$d(\, \ell t_{\text{mix}}(\varepsilon)\,) \leq \bar{d}(\, \ell t_{\text{mix}}(\varepsilon)\,) \leq \bar{d}(\, t_{\text{mix}}(\varepsilon)\,)^{\ell} \leq (2\varepsilon)^{\ell}. \tag{4.34}$$

In particular, taking $\varepsilon = 1/4$ above yields

$$d(\, \ell t_{\text{mix}}\,) \leq 2^{-\ell} \tag{4.35}$$

and

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \log_2 \varepsilon^{-1} \right\rceil t_{\text{mix}}. \tag{4.36}$$

Thus, although the choice of $1/4$ is arbitrary in the definition (4.33) of t_{mix}, a value of ε less than $1/2$ is needed to make the inequality $d(\, \ell t_{\text{mix}}(\varepsilon)\,) \leq (2\varepsilon)^{\ell}$ in (4.34) non-trivial and to achieve an inequality of the form (4.36).

4.6. Mixing and Time Reversal

For a distribution μ on a group G, the ***inverse distribution*** $\widehat{\mu}$ is defined by $\widehat{\mu}(g) := \mu(g^{-1})$ for all $g \in G$. Let P be the transition matrix of the random walk with increment distribution μ. Then the random walk with increment distribution $\widehat{\mu}$ is exactly the time reversal $\widehat{P}$ (defined in (1.33)) of P.

In Proposition 2.14 we noted that when $\widehat{\mu} = \mu$, the random walk on G with increment distribution μ is reversible, so that $P = \widehat{P}$. Even when μ is not a symmetric distribution, however, the forward and reversed walks must be at the same distance from stationarity, as we will find useful in analyzing card shuffling in Chapters 6 and 8.

LEMMA 4.13. *Let P be the transition matrix of a random walk on a group G with increment distribution μ and let $\widehat{P}$ be that of the walk on G with increment distribution $\widehat{\mu}$. Let π be the uniform distribution on G. Then for any $t \geq 0$*

$$\left\| P^t(\text{id}, \cdot) - \pi \right\|_{\text{TV}} = \left\| \widehat{P}^t(\text{id}, \cdot) - \pi \right\|_{\text{TV}}.$$

PROOF. Let $(X_t) = (\text{id}, X_1, \dots)$ be a Markov chain with transition matrix P and initial state id. We can write $X_k = g_1 g_2 \dots g_k$, where the random elements $g_1, g_2, \dots \in G$ are independent choices from the distribution μ. Similarly, let (Y_t)

time t :	1	0	1	0	0	1	1	1	0	0	0	0
time $t+1$:	1	0	1	0	0	1	1	1	0	0	0	0
time $t+2$:	1	0	1	0	0	1	1	1	0	0	0	0

FIGURE 4.3. The winning streak for $n = 5$. Here $X_t = 2$, $X_{t+1} = 3$, and $X_{t+2} = 0$.

time t :	1	0	1	0	0	1	1	1	0	0	0	0
time $t+1$:	1	0	1	0	0	1	1	1	0	0	0	0
time $t+2$:	1	0	1	0	0	1	1	1	0	0	0	0

FIGURE 4.4. The time reversal of the winning streak for $n = 5$. Here $\widehat{X}_t = 0$, $\widehat{X}_{t+1} = 3$, and $\widehat{X}_{t+2} = 2$.

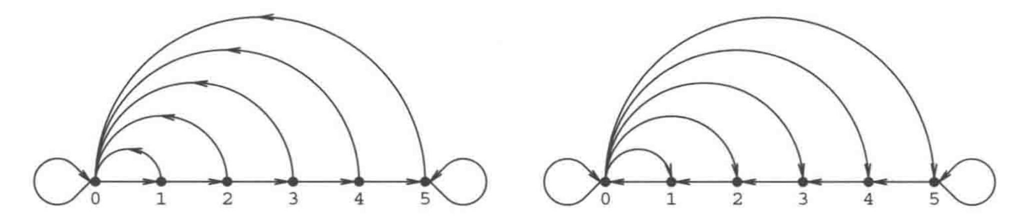


FIGURE 4.5. The underlying graphs of the transitions of (a) the winning streak chain for $n = 5$ and (b) its time reversal.

be a chain with transition matrix $\widehat{P}$, with increments $h_1, h_2, \dots \in G$ chosen independently from $\widehat{\mu}$. For any fixed elements $a_1, \dots, a_t \in G$,

$$\mathbf{P}\{g_1 = a_1, \dots, g_t = a_t\} = \mathbf{P}\{h_1 = a_t^{-1}, \dots, h_t = a_1^{-1}\},$$

by the definition of $\widehat{P}$. Summing over all strings such that $a_1 a_2 \dots a_t = a$ yields

$$P^t(\mathrm{id}, a) = \widehat{P}^t(\mathrm{id}, a^{-1}).$$

Hence

$$\sum_{a \in G} \left| P^t(\mathrm{id}, a) - |G|^{-1} \right| = \sum_{a \in G} \left| \widehat{P}^t(\mathrm{id}, a^{-1}) - |G|^{-1} \right| = \sum_{a \in G} \left| \widehat{P}^t(\mathrm{id}, a) - |G|^{-1} \right|$$

which together with Proposition 4.2 implies the desired result. ■

COROLLARY 4.14. *If t_{mix} is the mixing time of a random walk on a group and $\widehat{t_{\mathrm{mix}}}$ is the mixing time of the inverse walk, then $t_{\mathrm{mix}} = \widehat{t_{\mathrm{mix}}}$.*

EXAMPLE 4.15. It is also possible for reversing a Markov chain to significantly change the mixing time. The ***winning streak*** is an example. Here we bound the mixing time of its time reversal. The mixing time of the winning streak itself is discussed in Section 5.3.5.

Imagine a creature with bounded memory tossing a fair coin repeatedly and trying to track the length of the last run of heads. If there have been more than n heads in a row, the creature only remembers n of them. Hence the current state of our chain is the minimum of n and the length of the last run of heads.

Equivalently, consider a window of width n moving rightwards along an infinite string of independent fair bits, and let X_t be the length of the block of 1's starting at the right endpoint of the window. Then (X_t) is a Markov chain with state space $\{0, \dots, n\}$ and non-zero transitions given by

$$\begin{aligned} P(i,0) &= 1/2 \text{ for } 0 \le i \le n, \\ P(i,i+1) &= 1/2 \text{ for } 0 \le i < n, \\ P(n,n) &= 1/2. \end{aligned} \tag{4.37}$$

See Figures 4.3 and 4.5. It is straightforward to check that

$$\pi(i) = \begin{cases} 1/2^{i+1} & \text{if } i = 0, 1, \dots, n-1, \\ 1/2^n & \text{if } i = n \end{cases} \tag{4.38}$$

is stationary for P. It is also straightforward to check that the time reversal of P has non-zero entries

$$\begin{aligned} \widehat{P}(0,i) &= \pi(i) \text{ for } 0 \le i \le n, \\ \widehat{P}(i,i-1) &= 1 \text{ for } 1 \le i < n, \\ \widehat{P}(n,n) = P(n,n-1) &= 1/2. \end{aligned} \tag{4.39}$$

The coin-flip interpretation of the winning streak carries over to its time reversal. Imagine a window of width n moving *leftwards* along a string of independent random bits. Then the sequence of lengths $(\widehat{X}_t)$ of the rightmost block of 1's in the window is a version of the reverse winning streak chain. See Figures 4.4 and 4.5.

The time reversal of the mixing streak has the following remarkable property: after n steps, its distribution is exactly stationary, regardless of initial distribution. Why? Note first that if $\widehat{X}_t = 0$, then the distribution of $\widehat{X}_{t'}$ is stationary for all $t' > t$, since $\widehat{P}(0,\cdot) = \pi$. If $\widehat{X}_0 = k < n$, then the determined transitions force $X_k = 0$, so the chain is stationary for $t > k$. If $\widehat{X}_0 = n$, then the location of $\widehat{X_n}$ depends on the amount of time the chain spends at n before leaving. For $0 < k < n$, the chain has probability $1/2^k$ of holding $k-1$ times, then moving on the k-th turn. In this case $\widehat{X}_k = n-1$ and $\widehat{X}_n = k-1$. Also, $\widehat{P}^n(n,n) = 1/2^n$, so $\widehat{P}^n(n,\cdot) = \pi$. Finally, if the initial distribution is not concentrated at a single state, the distribution at time n is a mixture of the distributions from each possible starting state and is thus stationary.

For a lower bound, note that if the chain is started at n and leaves immediately, then at time $n-1$ it must be at state 1. Hence $\widehat{P}^{n-1}(n,1) = 1/2$, and the definition (4.1) of total variation distance implies that

$$d(n-1) \ge |\widehat{P}^{n-1}(n,1) - \pi(1)| = \frac{1}{4}.$$

We conclude that for the reverse winning streak chain, we have

$$t_{\mathrm{mix}}(\varepsilon) = n$$

for any positive $\varepsilon \le 1/4$.

4.7. Ergodic Theorem*

The idea of the ergodic theorem for Markov chains is that "time averages equal space averages".

If f is a real-valued function defined on Ω and μ is any probability distribution on Ω, then we define

$$E_\mu(f) = \sum_{x\in\Omega} f(x)\mu(x).$$

THEOREM 4.16 (Ergodic Theorem). *Let f be a real-valued function defined on Ω. If (X_t) is an irreducible Markov chain, then for any starting distribution μ,*

$$\mathbf{P}_\mu\left\{\lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = E_\pi(f)\right\} = 1. \tag{4.40}$$

PROOF. Suppose that the chain starts at x. Define $\tau_{x,0}^+ := 0$ and

$$\tau_{x,k}^+ := \min\{t > \tau_{x,(k-1)}^+ \,:\, X_t = x\}.$$

Since the chain "starts afresh" every time it visits x, the blocks $X_{\tau_{x,k}^+}, X_{\tau_{x,k}^+ + 1}, \ldots, X_{\tau_{x,(k+1)}^+ - 1}$ are independent of one another. Thus if

$$Y_k := \sum_{s=\tau_{x,(k-1)}^+}^{\tau_{x,k}^+ - 1} f(X_s),$$

then the sequence (Y_k) is i.i.d. If $S_t = \sum_{s=0}^{t-1} f(X_s)$, then $S_{\tau_{x,n}^+} = \sum_{k=1}^n Y_k$, and by the Strong Law of Large Numbers (Theorem A.8),

$$\mathbf{P}_x\left\{\lim_{n\to\infty}\frac{S_{\tau_{x,n}^+}}{n} = \mathbf{E}_x(Y_1)\right\} = 1.$$

Again by the Strong Law of Large Numbers, since $\tau_{x,n}^+ = \sum_{k=1}^n (\tau_{x,k}^+ - \tau_{x,(k-1)}^+)$, writing simply τ_x^+ for $\tau_{x,1}^+$,

$$\mathbf{P}_x\left\{\lim_{n\to\infty}\frac{\tau_{x,n}^+}{n} = \mathbf{E}_x(\tau_x^+)\right\} = 1.$$

Thus,

$$\mathbf{P}_x\left\{\lim_{n\to\infty}\frac{S_{\tau_{x,n}^+}}{\tau_{x,n}^+} = \frac{\mathbf{E}_x(Y_1)}{\mathbf{E}_x(\tau_x^+)}\right\} = 1. \tag{4.41}$$

Note that

$$\begin{aligned}\mathbf{E}_x(Y_1) &= \mathbf{E}_x\left(\sum_{s=0}^{\tau_x^+ - 1} f(X_s)\right) \\ &= \mathbf{E}_x\left(\sum_{y\in\Omega} f(y)\sum_{s=0}^{\tau_x^+ - 1}\mathbf{1}_{\{X_s=y\}}\right) = \sum_{y\in\Omega} f(y)\mathbf{E}_x\left(\sum_{s=0}^{\tau_x^+ - 1}\mathbf{1}_{\{X_s=y\}}\right).\end{aligned}$$

Using (1.25) shows that

$$\mathbf{E}_x(Y_1) = E_\pi(f)\mathbf{E}_x(\tau_x^+). \tag{4.42}$$

Putting together (4.41) and (4.42) shows that

$$\mathbf{P}_x\left\{\lim_{n\to\infty}\frac{S_{\tau^+_{x,n}}}{\tau^+_{x,n}} = E_\pi(f)\right\} = 1.$$

Exercise 4.2 shows that (4.40) holds when $\mu = \delta_x$, the probability distribution with unit mass at x. Averaging over the starting state completes the proof. ■

Taking $f(y) = \delta_x(y) = \mathbf{1}_{\{y=x\}}$ in Theorem 4.16 shows that

$$\mathbf{P}_\mu\left\{\lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1}\mathbf{1}_{\{X_s=x\}} = \pi(x)\right\} = 1,$$

so the asymptotic proportion of time the chain spends in state x equals $\pi(x)$.

Exercises

Exercise 4.1. Prove that

$$d(t) = \sup_\mu \left\|\mu P^t - \pi\right\|_{\mathrm{TV}},$$
$$\bar{d}(t) = \sup_{\mu,\nu} \left\|\mu P^t - \nu P^t\right\|_{\mathrm{TV}},$$

where μ and ν vary over probability distributions on a finite set Ω.

Exercise 4.2. Let (a_n) be a bounded sequence. If, for a sequence of integers (n_k) satisfying $\lim_{k\to\infty} n_k/n_{k+1} = 1$, we have

$$\lim_{k\to\infty}\frac{a_1+\cdots+a_{n_k}}{n_k} = a,$$

then

$$\lim_{n\to\infty}\frac{a_1+\cdots+a_n}{n} = a.$$

Exercise 4.3. Let P be the transition matrix of a Markov chain with state space Ω and let μ and ν be any two distributions on Ω. Prove that

$$\|\mu P - \nu P\|_{\mathrm{TV}} \le \|\mu - \nu\|_{\mathrm{TV}}.$$

(This in particular shows that $\left\|\mu P^{t+1} - \pi\right\|_{\mathrm{TV}} \le \|\mu P^t - \pi\|_{\mathrm{TV}}$, that is, advancing the chain can only move it closer to stationarity.)

Exercise 4.4. Let P be the transition matrix of a Markov chain with stationary distribution π. Prove that for any $t \ge 0$,

$$d(t+1) \le d(t),$$

where $d(t)$ is defined by (4.22).

Exercise 4.5. For $i = 1, \ldots, n$, let μ_i and ν_i be measures on Ω_i, and define measures μ and ν on $\prod_{i=1}^n \Omega_i$ by $\mu := \prod_{i=1}^n \mu_i$ and $\nu := \prod_{i=1}^n \nu_i$. Show that

$$\|\mu - \nu\|_{\mathrm{TV}} \le \sum_{i=1}^n \|\mu_i - \nu_i\|_{\mathrm{TV}}.$$

Notes

Our exposition of the Convergence Theorem follows Aldous and Diaconis (1986). Another approach is to study the eigenvalues of the transition matrix. See, for instance, Seneta (2006). Eigenvalues and eigenfunctions are often useful for bounding mixing times, particularly for reversible chains, and we will study them in Chapters 12 and 13. For convergence theorems for chains on infinite state spaces, see Chapter 21.

Aldous (1983b, Lemma 3.5) is a version of our Lemma 4.12 and Exercise 4.4. He says all these results "can probably be traced back to Doeblin."

The winning streak example is taken from Lovász and Winkler (1998).

We emphasize total variation distance, but mixing time can be defined using other distances. In particular, for $1 \leq p < \infty$, the $\ell^p(\pi)$ distance between μ and ν is defined as

$$\|\mu - \nu\|_p = \left[\sum_{x \in \Omega} \left|\frac{\mu(x)}{\pi(x)} - \frac{\nu(x)}{\pi(x)}\right|^p \pi(x)\right]^{1/p}.$$

The $\ell^\infty(\pi)$ distance is

$$\|\mu - \nu\|_\infty = \max_{x \in \Omega} \left|\frac{\mu(x)}{\pi(x)} - \frac{\nu(x)}{\pi(x)}\right|.$$

The separation distance, defined in Chapter 6, is often used.

The ***Hellinger distance*** d_H, defined as

$$d_H(\mu, \nu) := \sqrt{\sum_{x \in \Omega} \left(\sqrt{\mu(x)} - \sqrt{\nu(x)}\right)^2}, \tag{4.43}$$

behaves well on products (cf. Exercise 20.5). This distance is used in Section 20.4 to obtain a good bound on the mixing time for continuous product chains.

Further reading. Lovász (1993) gives the combinatorial view of mixing. Saloff-Coste (1997) and Montenegro and Tetali (2006) emphasize analytic tools. Aldous and Fill (1999) is indispensable. Other references include Sinclair (1993), Häggström (2002), Jerrum (2003), and, for an elementary account of the Convergence Theorem, Grinstead and Snell (1997, Chapter 11).

Complements. The result of Lemma 4.13 generalizes to transitive Markov chains, which we defined in Section 2.6.2.

LEMMA 4.17. *Let P be the transition matrix of a transitive Markov chain with state space Ω, let $\widehat{P}$ be its time reversal, and let π be the uniform distribution on Ω. Then*

$$\left\|\widehat{P}^t(x, \cdot) - \pi\right\|_{\mathrm{TV}} = \left\|P^t(x, \cdot) - \pi\right\|_{\mathrm{TV}}. \tag{4.44}$$

PROOF. Since our chain is transitive, for every $x, y \in \Omega$ there exists a bijection $\varphi_{(x,y)} : \Omega \to \Omega$ that carries x to y and preserves transition probabilities.

Now, for any $x, y \in \Omega$ and any t,

$$\sum_{z \in \Omega} \left|P^t(x, z) - |\Omega|^{-1}\right| = \sum_{z \in \Omega} \left|P^t(\varphi_{(x,y)}(x), \varphi_{(x,y)}(z)) - |\Omega|^{-1}\right| \tag{4.45}$$

$$= \sum_{z \in \Omega} \left|P^t(y, z) - |\Omega|^{-1}\right|. \tag{4.46}$$

Averaging both sides over y yields

$$\sum_{z\in\Omega} \left|P^t(x,z) - |\Omega|^{-1}\right| = \frac{1}{|\Omega|}\sum_{y\in\Omega}\sum_{z\in\Omega} \left|P^t(y,z) - |\Omega|^{-1}\right|. \tag{4.47}$$

Because π is uniform, we have $P(y,z) = \widehat{P}(z,y)$, and thus $P^t(y,z) = \widehat{P}^t(z,y)$. It follows that the right-hand side above is equal to

$$\frac{1}{|\Omega|}\sum_{y\in\Omega}\sum_{z\in\Omega} \left|\widehat{P}^t(z,y) - |\Omega|^{-1}\right| = \frac{1}{|\Omega|}\sum_{z\in\Omega}\sum_{y\in\Omega} \left|\widehat{P}^t(z,y) - |\Omega|^{-1}\right|. \tag{4.48}$$

By Exercise 2.8, $\widehat{P}$ is also transitive, so (4.47) holds with $\widehat{P}$ replacing P (and z and y interchanging roles). We conclude that

$$\sum_{z\in\Omega} \left|P^t(x,z) - |\Omega|^{-1}\right| = \sum_{y\in\Omega} \left|\widehat{P}^t(x,y) - |\Omega|^{-1}\right|. \tag{4.49}$$

Dividing by 2 and applying Proposition 4.2 completes the proof. ■

REMARK 4.18. The proof of Lemma 4.13 established an exact correspondence between forward and reversed trajectories, while that of Lemma 4.17 relied on averaging over the state space.

CHAPTER 5

Coupling

5.1. Definition

As we defined in Section 4.1, a coupling of two probability distributions μ and ν is a pair of random variables (X, Y), defined on the same probability space, such that the marginal distribution of X is μ and the marginal distribution of Y is ν.

Couplings are useful because a comparison between distributions is reduced to a comparison between random variables. Proposition 4.7 characterized $\|\mu - \nu\|_{\mathrm{TV}}$ as the minimum, over all couplings (X, Y) of μ and ν, of the probability that X and Y are different. This provides an effective method of obtaining upper bounds on the distance.

In this chapter, we will extract more information by coupling not only pairs of distributions, but entire Markov chain trajectories. Here is a simple initial example.

EXAMPLE 5.1. A simple random walk on the segment $\{0, 1, \ldots, n\}$ is a Markov chain which moves either up or down at each move with equal probability. If the walk attempts to move outside the interval when at a boundary point, it stays put. It is intuitively clear that $P^t(x, n) \leq P^t(y, n)$ whenever $x \leq y$, as this says that the chance of being at the "top" value n after t steps does not decrease as you increase the height of the starting position.

A simple proof uses a coupling of the distributions $P^t(x, \cdot)$ and $P^t(y, \cdot)$. Let $\Delta_1, \Delta_2, \ldots$ be a sequence of i.i.d. (that is, independent and identically distributed) $\{-1, 1\}$-valued random variables with zero mean, so each Δ_i is equally likely to be $+1$ as -1. We will define together two random walks on $\{0, 1, \ldots, n\}$: the walk (X_t) starts at x, while the walk (Y_t) starts at y.

We use the same rule for moving in both chains (X_t) and (Y_t): if $\Delta_t = +1$, move the chain up if possible, and if $\Delta_t = -1$, move the chain down if possible. Hence the chains move in step, although they are started at different heights. Once the two chains meet (necessarily either at 0 or n), they stay together thereafter.

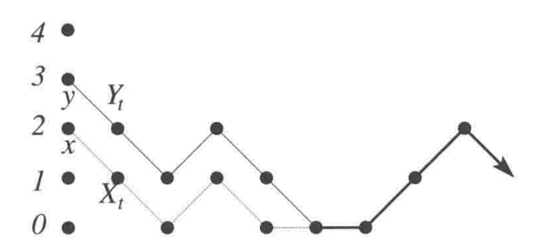


FIGURE 5.1. Coupled random walks on $\{0, 1, 2, 3, 4\}$. The walks stay together after meeting.

Clearly the distribution of X_t is $P^t(x, \cdot)$, and the distribution of Y_t is $P^t(y, \cdot)$. Importantly, X_t and Y_t are defined on the same underlying probability space, as both chains use the sequence (Δ_t) to determine their moves.

It is clear that if $x \leq y$, then $X_t \leq Y_t$ for all t. In particular, if $X_t = n$, the top state, then it must be that $Y_t = n$ also. From this we can conclude that

$$P^t(x, n) = \mathbf{P}\{X_t = n\} \leq \mathbf{P}\{Y_t = n\} = P^t(y, n). \tag{5.1}$$

This argument shows the power of coupling. We were able to couple together the two chains in such a way that $X_t \leq Y_t$ always, and from this fact about the random variables we could easily read off information about the distributions.

In the rest of this chapter, we will see how building two simultaneous copies of a Markov chain using a common source of randomness, as we did in the previous example, can be useful for getting bounds on the distance to stationarity.

We define a ***coupling of Markov chains*** with transition matrix P to be a process $(X_t, Y_t)_{t=0}^{\infty}$ with the property that both (X_t) and (Y_t) are Markov chains with transition matrix P, although the two chains may possibly have different starting distributions.

Any coupling of Markov chains with transition matrix P can be modified so that the two chains stay together at all times after their first simultaneous visit to a single state—more precisely, so that

$$\text{if } X_s = Y_s, \text{ then } X_t = Y_t \text{ for } t \geq s. \tag{5.2}$$

To construct a coupling satisfying (5.2), simply run the chains according to the original coupling until they meet; then run them together.

NOTATION. If (X_t) and (Y_t) are coupled Markov chains with $X_0 = x$ and $Y_0 = y$, then we will often write $\mathbf{P}_{x,y}$ for the probability on the space where (X_t) and (Y_t) are both defined.

5.2. Bounding Total Variation Distance

As usual, we will fix an irreducible transition matrix P on the state space Ω and write π for its stationary distribution. The following is the key tool used in this chapter.

THEOREM 5.2. *Let $\{(X_t, Y_t)\}$ be a coupling satisfying* (5.2) *for which $X_0 = x$ and $Y_0 = y$. Let τ_{couple} be the first time the chains meet:*

$$\tau_{\text{couple}} := \min\{t : X_t = Y_t\}. \tag{5.3}$$

Then

$$\left\|P^t(x, \cdot) - P^t(y, \cdot)\right\|_{\text{TV}} \leq \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t\}. \tag{5.4}$$

PROOF. Notice that $P^t(x, z) = \mathbf{P}_{x,y}\{X_t = z\}$ and $P^t(y, z) = \mathbf{P}_{x,y}\{Y_t = z\}$. Consequently, (X_t, Y_t) is a coupling of $P^t(x, \cdot)$ with $P^t(y, \cdot)$, whence Proposition 4.7 implies that

$$\left\|P^t(x, \cdot) - P^t(y, \cdot)\right\|_{\text{TV}} \leq \mathbf{P}_{x,y}\{X_t \neq Y_t\}. \tag{5.5}$$

By (5.2), $\mathbf{P}_{x,y}\{X_t \neq Y_t\} = \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t\}$, which with (5.5) establishes (5.4). ■

Combining Theorem 5.2 with Lemma 4.11 proves the following:

COROLLARY 5.3. *Suppose that for each pair of states $x, y \in \Omega$ there is a coupling (X_t, Y_t) with $X_0 = x$ and $Y_0 = y$. For each such coupling, let τ_{couple} be the first time the chains meet, as defined in (5.3). Then*

$$d(t) \le \max_{x,y\in\Omega} \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t\}.$$

Given a Markov chain on Ω with transition matrix P, a ***Markovian coupling*** of P is a Markov chain with state space $\Omega \times \Omega$ whose transition matrix Q satisfies

(i) for all x, y, x' we have $\sum_{y'} Q((x,y),(x',y')) = P(x,x')$ and
(ii) for all x, y, y' we have $\sum_{x'} Q((x,y),(x',y')) = P(y,y')$.

Clearly any Markovian coupling is indeed a coupling of Markov chains, as we defined in Section 5.1.

REMARK 5.4. All couplings used in this book will be Markovian.

5.3. Examples

5.3.1. Random walk on the cycle. We defined random walk on the n-cycle in Example 1.4. The underlying graph of this walk, $\mathbb{Z}_n$, has vertex set $\{1, 2, \dots, n\}$ and edges between j and k whenever $j \equiv k \pm 1 \mod n$. See Figure 1.3.

We consider the lazy walk, which remains in its current position with probability $1/2$, moves clockwise with probability $1/4$, and moves counterclockwise with probability $1/4$.

We construct a coupling (X_t, Y_t) of two particles performing lazy walks on $\mathbb{Z}_n$, one started from x and the other started from y. In this coupling, the two particles will never move simultaneously, ensuring that they will not jump over one another when they come to within unit distance. At each move, a fair coin is tossed, independent of all previous tosses. If heads, the chain (X_t) moves one step, the direction of which is determined by another fair coin toss, again independent of all other previous tosses. If tails, the chain (Y_t) moves one step, also determined by an independent fair coin toss. Once the two particles collide, thereafter they make identical moves. Let D_t be the clockwise distance between the two particles. Note that D_t is a simple random walk on the interior vertices of $\{0, 1, 2, \dots, n\}$ and gets absorbed at either 0 or n. By Proposition 2.1, if $\tau = \min\{t \ge 0 : D_t \in \{0, n\}\}$, then $\mathbf{E}_{x,y}(\tau) = k(n-k)$, where k is the clockwise distance between x and y. Since $\tau = \tau_{\text{couple}}$, by Corollary 5.3,

$$d(t) \le \max_{x,y\in\mathbb{Z}_n} \mathbf{P}_{x,y}\{\tau > t\} \le \frac{\max_{x,y} \mathbf{E}_{x,y}(\tau)}{t} \le \frac{n^2}{4t}.$$

The right-hand side equals $1/4$ for $t = n^2$, whence $t_{\text{mix}} \le n^2$.

In Section 7.4.1, it is shown that $t_{\text{mix}} \ge c_1 n^2$ for a constant c_1.

5.3.2. Random walk on the torus. The *d-dimensional torus* is graph whose vertex set is the Cartesian product

$$\mathbb{Z}_n^d = \underbrace{\mathbb{Z}_n \times \cdots \times \mathbb{Z}_n}_{d \text{ times}}.$$

Vertices $\boldsymbol{x} = (x^1, \dots, x^d)$ and $\boldsymbol{y} = (y^1, y^2, \dots, y^d)$ are neighbors in $\mathbb{Z}_n^d$ if for some $j \in \{1, 2, \dots, n\}$, we have $x^i = y^i$ for all $i \ne j$ and $x^j \equiv y^j \pm 1 \mod n$. See Figure 5.2 for an example.

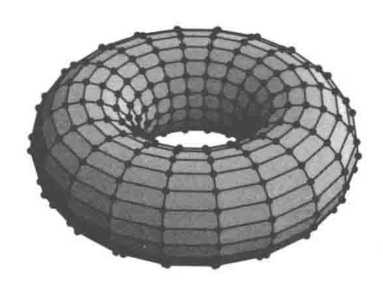


FIGURE 5.2. The 2-torus $\mathbb{Z}_{20}^2$.

When n is even, the graph $\mathbb{Z}_n^d$ is bipartite and the associated random walk is periodic. To avoid this complication, we consider the lazy random walk on $\mathbb{Z}_n^d$, defined in Section 1.3. This walk remains at its current position with probability $1/2$ at each move.

We now use coupling to bound the mixing time of the lazy random walk on $\mathbb{Z}_n^d$.

THEOREM 5.5. *For the lazy random walk on the d-dimension torus* $\mathbb{Z}_n^d$,

$$t_{\text{mix}}(\varepsilon) \leq c(d)n^2 \log_2(\varepsilon^{-1}), \tag{5.6}$$

where $c(d)$ is a constant depending on the dimension d.

PROOF. In order to apply Corollary 5.3 to prove this theorem, we construct a coupling for each pair $(\boldsymbol{x}, \boldsymbol{y})$ of starting states and bound the expected value of the coupling time $\tau_{\text{couple}} = \tau_{\boldsymbol{x},\boldsymbol{y}}$.

To couple together a random walk $(\boldsymbol{X}_t)$ started at $\boldsymbol{x}$ with a random walk $(\boldsymbol{Y}_t)$ started at $\boldsymbol{y}$, first pick one of the d coordinates at random. If the positions of the two walks agree in the chosen coordinate, we move both of the walks by $+1$, -1, or 0 in that coordinate, with probabilities $1/4$, $1/4$ and $1/2$, respectively. If the positions of the two walks differ in the chosen coordinate, we randomly choose one of the chains to move, leaving the other fixed. We then move the selected walk by $+1$ or -1 in the chosen coordinate, with the sign determined by a fair coin toss.

Let $\boldsymbol{X}_t = (X_t^1, \ldots, X_t^d)$ and $\boldsymbol{Y}_t = (Y_t^1, \ldots, Y_t^d)$, and let

$$\tau_i := \min\{t \geq 0 : X_t^i = Y_t^i\}$$

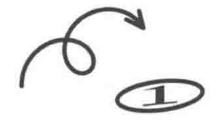
be the time required for the chains to agree in coordinate i.

The clockwise difference between X_t^i and Y_t^i, viewed at the times when coordinate i is selected, behaves just as the coupling of the lazy walk on the cycle $\mathbb{Z}_n$ discussed above. Thus, the expected number of moves in coordinate i needed to make the two chains agree on that coordinate is not more than $n^2/4$.

Since coordinate i is selected with probability $1/d$ at each move, there is a geometric waiting time between moves with expectation d. Exercise 5.3 implies that

$$\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}}(\tau_i) \leq \frac{dn^2}{4}. \tag{5.7}$$

Copy 1:	0	0	1	1	0	1	0	0	1	1
Copy 2:	0	1	1	0	0	0	1	0	1	0



Copy 1:	0	0	1	1	0	1	0	0	1	1
Copy 2:	0	1	1	0	0	1	1	0	1	0

FIGURE 5.3. One step in two coupled lazy walks on the hypercube. First, choose a coordinate to update—here, the sixth. Then, flip a 0/1 coin and use the result to update the chosen coordinate to the same value in both walks.

The coupling time we are interested in is $\tau_{\text{couple}} = \max_{1 \le i \le d} \tau_i$, and we can bound the maximum by a sum to get

$$\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}}(\tau_{\text{couple}}) \le \frac{d^2 n^2}{4}. \tag{5.8}$$

This bound is independent of the starting states, and we can use Markov's inequality to show that

$$\mathbf{P}_{\boldsymbol{x},\boldsymbol{y}}\{\tau_{\text{couple}} > t\} \le \frac{\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}}(\tau_{\text{couple}})}{t} \le \frac{1}{t}\frac{d^2 n^2}{4}. \tag{5.9}$$

Taking $t_0 = d^2 n^2$ shows that $d(t_0) \le 1/4$, and so $t_{\text{mix}} \le d^2 n^2$. By (4.36),

$$t_{\text{mix}}(\varepsilon) \le d^2 n^2 \left\lceil \log(\varepsilon^{-1}) \right\rceil,$$

and we have proved the theorem. ■

Exercise 5.4 shows that the bound on $c(d)$ can be improved.

5.3.3. Random walk on the hypercube. The simple random walk on the hypercube $\{0,1\}^n$ was defined in Section 2.3: this is the simple walker on the graph having vertex set $\{0,1\}^n$, the binary words of length n, and with edges connecting words differing in exactly one letter. (Note that this graph is also the torus $\mathbb{Z}_2^n$.)

To avoid periodicity, we study the lazy chain: at each time step, the walker remains at her current position with probability $1/2$ and with probability $1/2$ moves to a position chosen uniformly at random among all neighboring vertices.

As remarked in Section 2.3, a convenient way to generate the lazy walk is as follows: pick one of the n coordinates uniformly at random, and *refresh* the bit at this coordinate with a random fair bit (one which equals 0 or 1 each with probability $1/2$).

This algorithm for running the walk leads to the following coupling of two walks with possibly different starting positions: first, pick among the n coordinates uniformly at random; suppose that coordinate i is selected. *In both walks*, replace the bit at coordinate i *with the same* random fair bit. (See Figure 5.3.) From this time onwards, both walks will agree in the i-th coordinate. A moment's thought reveals that individually each of the walks is indeed a lazy random walk on the hypercube.

If τ is the first time when all of the coordinates have been selected at least once, then the two walkers agree with each other from time τ onwards. (If the

initial states agree in some coordinates, the first time the walkers agree could be strictly before τ.) The distribution of τ is exactly the same as the coupon collector random variable studied in Section 2.2. Using Corollary 5.3, together with the bound on the tail of τ given in Proposition 2.4, shows that

$$d(n \log n + cn) \leq \mathbf{P}\{\tau > n \log n + cn\} \leq e^{-c}.$$

It is immediate from the above that

$$t_{\text{mix}}(\varepsilon) \leq n \log n + \log(1/\varepsilon) n. \tag{5.10}$$

Simply, $t_{\text{mix}} = O(n \log n)$. The bound in (5.10) is off by a factor of two and will be sharpened in Section 18.2.2 via a more sophisticated coupling.

5.3.4. Random walk on a finite binary tree. Since trees will appear in several examples in the sequel, we collect some definitions here. A ***tree*** is a connected graph with no cycles. (See Exercise 1.3 and Exercise 1.4.) A ***rooted*** tree has a distinguished vertex, called the ***root***. The ***depth*** of a vertex v is its graph distance to the root. A ***level*** of the tree consists of all vertices at the same depth. The ***children*** of v are the neighbors of v with depth larger than v. A ***leaf*** is a vertex with degree one.

A ***rooted finite b-ary tree of depth*** k, denoted by $T_{b,k}$, is a tree with a distinguished vertex v_0, the root, such that

- v_0 has degree b,
- every vertex at distance j from the root, where $1 \leq j \leq k-1$, has degree $b+1$,
- the vertices at distance k from v_0 are leaves.

There are $n = (b^{k+1} - 1)/(b-1)$ vertices in $T_{b,k}$.

In this example, we consider the lazy random walk on the finite ***binary tree*** $T_{2,k}$; this walk remains at its current position with probability $1/2$.

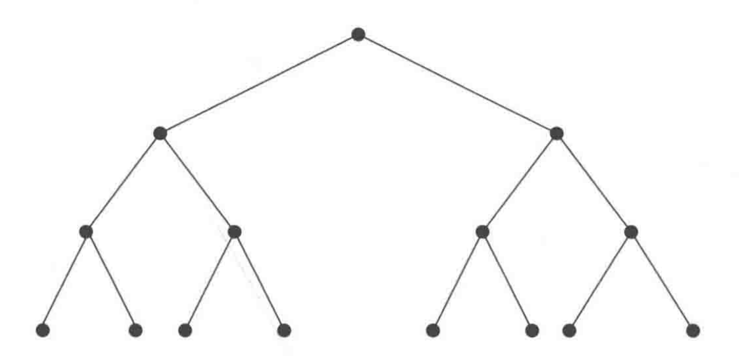


FIGURE 5.4. A binary tree of height 3.

Consider the following coupling (X_t, Y_t) of two lazy random walks, started from states x_0 and y_0 on the tree. Assume, without loss of generality, that x_0 is at least as close to the root as y_0. At each move, toss a fair coin to decide which of the two chains moves: if heads, $Y_{t+1} = Y_t$, while X_{t+1} is chosen from the neighbors of X_t uniformly at random. If the coin toss is tails, then $X_{t+1} = X_t$, and Y_{t+1} is chosen from the neighbors of Y_t uniformly at random. Run the two chains according to this rule until the first time they are at the same level of the tree. Once the two

chains are at the same level, change the coupling to the following updating rule: let X_t evolve as a lazy random walk, and couple Y_t to X_t so that Y_t moves closer to (further from) the root if and only if X_t moves closer to (further from) the root, respectively. Let B be the set of leaves. Observe that if (X_t) has first visited B and then visited the root, it must have coupled by this time. The expected value of this time is less than the expected ***commute time*** τ from the root to B, the time it takes starting from the root to first visit the set B and then return to the root. It will be shown in Example 10.12 that $\mathbf{E}(\tau) \leq 4n$. Thus, if τ_{couple} is the time when the two particles meet, we have $\mathbf{P}_{x,y}\{\tau_{\text{couple}} > t\} \leq \frac{4n}{t}$. We conclude that $t_{\text{mix}} \leq 16n$.

5.3.5. The winning streak. Recall the winning streak chain, defined in Example 4.15. For any initial values $a, b \in \{0, \ldots, n\}$, let x and y be bitstrings of length n whose ending block of 1's have length exactly a and b, respectively. Then we can couple copies of the winning streak chain started at a and at b by appending the same uniform random bits to the ends of x and y, then counting the number of terminal 1's in the resulting window.

As soon as one of the added bits is 0, both chains fall into state 0, and they remain coupled thereafter. Hence

$$\mathbf{P}\{\tau_{\text{couple}} > t\} \leq 2^{-t}$$

and Corollary 5.3 gives

$$d(t) \leq 2^{-t}.$$

By the definition (4.32) of mixing time, we have

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \log_2 \left(\frac{1}{\varepsilon} \right) \right\rceil,$$

which depends only on ε, and not on n. In particular, $t_{\text{mix}} \leq 2$ for all n. (In Example 4.15 we showed that the mixing time of the reversed winning streak was of order $O(n)$.)

5.3.6. Distance between $P^t(x, \cdot)$ and $P^{t+1}(x, \cdot)$.

PROPOSITION 5.6. *Let Q be an irreducible transition matrix and consider the lazy chain with transition matrix $P = (Q + I)/2$. The distributions at time t and $t + 1$ satisfy*

$$\left\| P^t(x, \cdot) - P^{t+1}(x, \cdot) \right\|_{\text{TV}} \leq \frac{12}{\sqrt{t}}. \tag{5.11}$$

PROOF. We construct two Markov chains, (X_t) and (Y_t), both with transition matrix P and both started at x, such that

$$\mathbf{P}\{X_t \neq Y_{t+1}\} \leq 12/\sqrt{t}. \tag{5.12}$$

Since the distribution of X_t is $P^t(x, \cdot)$ and the distribution of Y_{t+1} is $P^{t+1}(x, \cdot)$, the inequality in (5.12) along with Proposition 4.7 implies (5.11).

Let $(Z_t)_{t=1}^{\infty}$ be a Markov chain with transition matrix Q started from x, and let $(W_t)_{t=1}^{\infty}$ be an i.i.d. sequence of unbiased $\{0, 1\}$-valued random variables, independent of (Z_t). Define $N_t := \sum_{s=1}^{t} W_t$ and $Y_t := Z_{N_t}$. For $t \geq 1$, define

$$X_t := \begin{cases} Z_{t-(N_{t+1}-W_1)} & \text{if } X_{t-1} \neq Y_t, \\ Y_{t+1} & \text{if } X_{t-1} = Y_t. \end{cases}$$

The reader should convince himself that both (X_t) and (Y_t) are Markov chains with transition matrix P.

Let $\tau = \min\{t \geq 0 : X_t = Y_{t+1}\}$. If $W_1 = 0$, then $Y_1 = x = X_0$ and $\tau = 0$. If $W_1 = 1$, then

$$\tau \leq \min\{t \geq 0 : N_{t+1} = t - (N_{t+1} - W_1)\}.$$

Observe that on the event $\{W_1 = 1\}$, the equality $N_{t+1} = t - (N_{t+1} - W_1)$ holds if and only if $2(N_{t+1} - W_1) - t = -1$. Therefore, τ is stochastically bounded by the first time a simple random walk hits -1. By Theorem 2.17, $\mathbf{P}\{\tau > t\} \leq 12/\sqrt{t}$. This establishes the inequality in (5.12), finishing the proof. ■

5.4. Grand Couplings

It can be useful to construct simultaneously, using a common source of randomness, Markov chains started from each state in Ω. That is, we want to construct a collection of random variables $\{X_t^x : x \in \Omega, t = 0, 1, 2, \ldots\}$ such that for each $x \in \Omega$, the sequence $(X_t^x)_{t=0}^\infty$ is a Markov chain with transition matrix P started from x. We call such a collection a ***grand coupling***.

The random mapping representation of a chain, discussed in Section 1.2, can be used to construct a grand coupling. Let $f : \Omega \times \Lambda \to \mathbb{R}$ be a function and Z a Λ-valued random variable such that $P(x, y) = \mathbf{P}\{f(x, Z) = y\}$. Proposition 1.5 guarantees that such an (f, Z) pair exists. Let $Z_1, Z_2, \ldots$ be an i.i.d. sequence, each with the same distribution as Z, and define

$$X_0^x = x, \quad X_t^x = f(X_{t-1}^x, Z_t) \text{ for } t \geq 1. \tag{5.13}$$

Since each of $(X_t^x)_{t=0}^\infty$ is a Markov chain started from x with transition matrix P, this yields a grand coupling. We emphasize that the chains $(X_t^x)_{t=0}^\infty$ all live on the same probability space, each being determined by the same sequence of random variables $(Z_t)_{t=0}^\infty$.

5.4.1. Random colorings.

Random proper colorings of a graph were introduced in Section 3.3.1. For a graph G with vertex set V, let Ω be the set of proper colorings of G, and let π be the uniform distribution on Ω. In Example 3.5, the Metropolis chain for π was introduced. A transition for this chain is made by first selecting a vertex v uniformly from V and then selecting a color k uniformly from $\{1, 2, \ldots, q\}$. If placing color k at vertex v is permissible (that is, if no neighbor of v has color k), then vertex v is assigned color k. Otherwise, no transition is made.

Note that in fact this transition rule can be defined on the space $\tilde{\Omega}$ of all (not necessarily proper) colorings, and the grand coupling can be defined simultaneously for all colorings in $\tilde{\Omega}$.

Using grand couplings, we can prove the following theorem:

THEOREM 5.7. *Let G be a graph with n vertices and maximal degree Δ. For the Metropolis chain on proper colorings of G, if $q > 3\Delta$ and $c_{\text{met}}(\Delta, q) := 1 - (3\Delta/q)$, then*

$$t_{\text{mix}}(\varepsilon) \leq c_{\text{met}}(\Delta, q)^{-1} n \left[\log n + \log(1/\varepsilon)\right]. \tag{5.14}$$

In Chapter 14 we show that for Glauber dynamics on proper colorings (see Section 3.3 for the definition of this chain), if $q > 2\Delta$, then the mixing time is of order $n \log n$.

PROOF. Just as for a single Metropolis chain on colorings, the grand coupling at each move generates a single vertex and color pair (v, k), uniformly at random from $V \times \{1, \ldots, q\}$ and independent of the past. For each $x \in \tilde{\Omega}$, the coloring X_t^x is updated by attempting to re-color vertex v with color k, accepting the update if and only if the proposed new color is different from the colors at vertices neighboring v. (If a re-coloring is not accepted, the chain X_t^x remains in its current state.) The essential point is that the same vertex and color pair is used for all the chains (X_t^x).

For two colorings $x, y \in \tilde{\Omega}$, define

$$\rho(x, y) := \sum_{v \in V} \mathbf{1}_{\{x(v) \neq y(v)\}}$$

to be the number of vertices where x and y disagree, and note that ρ is a metric on $\tilde{\Omega}$.

Suppose $\rho(x, y) = 1$, so that x and y agree everywhere except at vertex v_0. Write $\mathcal{N}$ for the set of colors appearing among the neighbors of v_0 in x, which is the same as the set of colors appearing among the neighbors of v_0 in y. Recall that v represents a random sample from V, and k a random sample from $\{1, 2, \ldots, q\}$, independent of v. We consider the distance after updating x and y in one step of the grand coupling, that is, $\rho(X_1^x, X_1^y)$.

This distance goes to zero if and only if the vertex v_0 is selected for updating and the color proposed is not in $\mathcal{N}$. This occurs with probability

$$\mathbf{P}\{\rho(X_1^x, X_1^y) = 0\} = \left(\frac{1}{n}\right)\left(\frac{q - |\mathcal{N}|}{q}\right) \geq \frac{q - \Delta}{nq}, \tag{5.15}$$

where Δ denotes the maximum vertex degree in the graph.

For a vertex w which is a neighbor of v_0, note that the set of colors among the neighbors of w different from v_0 are the same in the colorings x and y. Suppose that neither $x(v_0)$ nor $y(v_0)$ belong to this set of colors. In this case, if w is the vertex selected for updating and the color $x(v_0)$ is proposed, then configuration y will be updated at w (to the color $x(v_0)$), while configuration x will not be updated. See Figure 5.5. This will cause the number of disagreements between x and y to increase to two. Similarly, the disagreements will increase if w is selected and the color $y(v_0)$ is proposed. These are the only scenarios leading to $\rho(X_1^x, X_1^y) = 2$, and we conclude that

$$\mathbf{P}\{\rho(X_1^x, X_1^y) = 2\} \leq \left(\frac{\Delta}{n}\right)\left(\frac{2}{q}\right). \tag{5.16}$$

Using the bounds (5.15) and (5.16),

$$\mathbf{E}\left(\rho(X_1^x, X_1^y) - 1\right) \leq \frac{2\Delta}{nq} - \frac{(q - \Delta)}{nq} = \frac{3\Delta - q}{nq},$$

and so

$$\mathbf{E}\left(\rho(X_1^x, X_1^y)\right) \leq 1 - \frac{q - 3\Delta}{nq}.$$

If $q > 3\Delta$, then $c_{\text{met}}(\Delta, q) = 1 - (3\Delta/q) > 0$ and

$$\mathbf{E}\left(\rho(X_1^x, X_1^y)\right) \leq 1 - \frac{c_{\text{met}}(\Delta, q)}{n} < 1. \tag{5.17}$$

Now, suppose that x and y are colorings with $\rho(x, y) = r$. There are colorings $x_0 = x, x_1, \ldots, x_r = y$ such that $\rho(x_k, x_{k-1}) = 1$. Since ρ is a metric and the

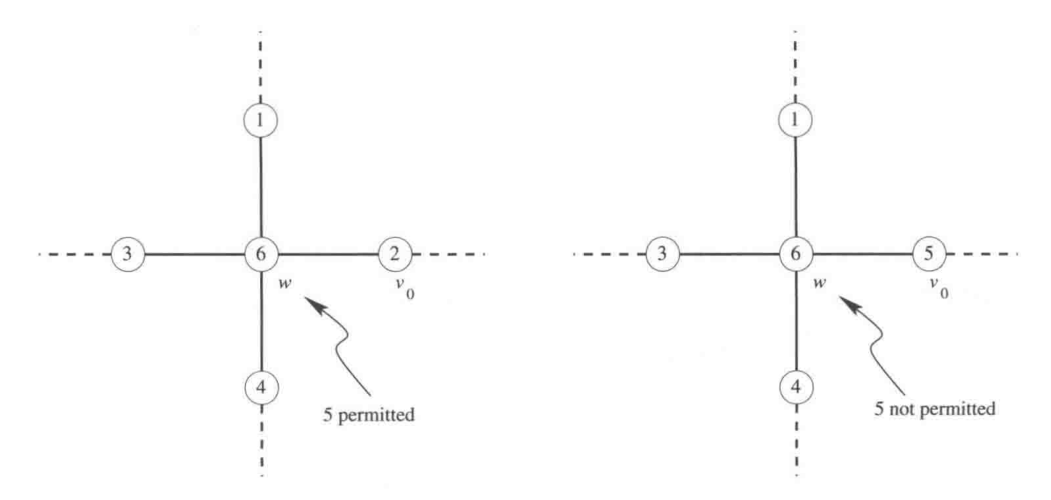


FIGURE 5.5. Two colorings which disagree only at v_0. The one on the left can be updated with the color 5 at a neighbor of w of v_0, while the one on the right cannot be updated with a 5 at w. If vertex w is selected for updating and color 5 is proposed, the two configurations will disagree at both v_0 and w.

inequality (5.17) can be applied to each of $\mathbf{E}\left(\rho(X_1^{x_k}, X_1^{x_{k-1}})\right)$,

$$\begin{aligned}\mathbf{E}\left(\rho(X_1^x, X_1^y)\right) &\leq \sum_{k=1}^{r} \mathbf{E}\left(\rho(X_1^{x_k}, X_1^{x_{k-1}})\right) \\ &\leq r\left(1 - \frac{c_{\text{met}}(\Delta, q)}{n}\right) = \rho(x,y)\left(1 - \frac{c_{\text{met}}(\Delta, q)}{n}\right).\end{aligned}$$

Conditional on the event that $X_{t-1}^x = x_{t-1}$ and $X_{t-1}^y = y_{t-1}$, the random vector (X_t^x, X_t^y) has the same distribution as $(X_1^{x_{t-1}}, X_1^{y_{t-1}})$. Hence,

$$\begin{aligned}\mathbf{E}\left(\rho(X_t^x, X_t^y) \mid X_{t-1}^x = x_{t-1}, X_{t-1}^y = y_{t-1}\right) &= \mathbf{E}\left(\rho(X_1^{x_{t-1}}, X_1^{y_{t-1}})\right) \\ &\leq \rho(x_{t-1}, y_{t-1})\left(1 - \frac{c_{\text{met}}(\Delta, q)}{n}\right).\end{aligned}$$

Taking an expectation over (X_{t-1}^x, X_{t-1}^y) shows that

$$\mathbf{E}\left(\rho(X_t^x, X_t^y)\right) \leq \mathbf{E}\left(\rho(X_{t-1}^x, X_{t-1}^y)\right)\left(1 - \frac{c_{\text{met}}(\Delta, q)}{n}\right).$$

Iterating the above inequality shows that

$$\mathbf{E}\left(\rho(X_t^x, X_t^y)\right) \leq \rho(x,y)\left(1 - \frac{c_{\text{met}}(\Delta, q)}{n}\right)^t.$$

Moreover, by Markov's inequality, since $\rho(x,y) \geq 1$ when $x \neq y$,

$$\begin{aligned}\mathbf{P}\{X_t^x \neq X_t^y\} &= \mathbf{P}\{\rho(X_t^x, X_t^y) \geq 1\} \\ &\leq \rho(x,y)\left(1 - \frac{c_{\text{met}}(\Delta, q)}{n}\right)^t \leq n e^{-t(c_{\text{met}}(\Delta, q)/n)}.\end{aligned}$$

Since the above holds for all colorings $x, y \in \tilde{\Omega}$, in particular it holds for all proper colorings $x, y \in \Omega$. By Corollary 5.3 and the above inequality, $d(t) \leq ne^{-t(c_{\text{met}}(\Delta,q)/n)}$, whence if

$$t > c_{\text{met}}(\Delta, q)^{-1} n \left[\log n + \log(1/\varepsilon)\right],$$

then $d(t) \leq \varepsilon$. This establishes (5.14). ∎

5.4.2. Hardcore model. The hardcore model with fugacity λ was introduced in Section 3.3.4. We use a grand coupling to show that if λ is small enough, the Glauber dynamics has a mixing time of the order $n \log n$.

THEOREM 5.8. *Let $c_H(\lambda) := [1 + \lambda(1 - \Delta)]/(1 + \lambda)$. For the Glauber dynamics for the hardcore model on a graph with maximum degree Δ and n vertices, if $\lambda < (\Delta - 1)^{-1}$, then*

$$t_{\text{mix}}(\varepsilon) \leq \frac{n}{c_H(\lambda)}[\log n + \log(1/\varepsilon)].$$

PROOF. We again use the grand coupling which is run as follows: a vertex v is selected uniformly at random, and a coin with probability $\lambda/(1 + \lambda)$ of heads is tossed, independently of the choice of v. Each hardcore configuration x is updated using v and the result of the coin toss. If the coin is tails, any particle present at v in x is removed. If the coin is heads and all neighbors of v are unoccupied in the configuration x, then a particle is placed at v.

We let $\rho(x, y) = \sum_{v \in V} \mathbf{1}_{\{x(v) \neq y(v)\}}$ be the number of sites where x and y disagree. Suppose that x and y satisfy $\rho(x, y) = 1$, so that the two configurations differ only at v_0. Without loss of generality, assume that $x(v_0) = 1$ and $y(v_0) = 0$.

If vertex v_0 is selected, then $\rho(X_1^x, X_1^y) = 0$, since the neighbors of v_0 agree in both x and y so the same action will be taken for the two configurations.

Let w be a neighbor of v_0. If none of the neighbors of w different from v_0 are occupied (these sites have the same status in x and y) and the coin toss is heads, then x and y will be updated differently. Indeed, it will be possible to place a particle at w in y, but not in x. This is the only case in which a new disagreement between x and y can be introduced.

Therefore,

$$\mathbf{E}\left(\rho(X_1^x, X_1^y)\right) \leq 1 - \frac{1}{n} + \frac{\Delta}{n}\frac{\lambda}{1+\lambda} = 1 - \frac{1}{n}\left[\frac{1 - \lambda(\Delta - 1)}{1 + \lambda}\right].$$

If $\lambda < (\Delta - 1)^{-1}$, then $c_H(\lambda) > 0$ and

$$\mathbf{E}\left(\rho(X_1^x, X_1^y)\right) \leq 1 - \frac{c_H(\lambda)}{n} \leq e^{-c_H(\lambda)/n}.$$

The remainder of the theorem follows exactly the same argument as is used at the end of Theorem 5.7. ∎

Exercises

EXERCISE 5.1. A mild generalization of Theorem 5.2 can be used to give an alternative proof of the Convergence Theorem.

(a) Show that when (X_t, Y_t) is a coupling satisfying (5.2) for which $X_0 \sim \mu$ and $Y_0 \sim \nu$, then

$$\left\|\mu P^t - \nu P^t\right\|_{\text{TV}} \leq \mathbf{P}\{\tau_{\text{couple}} > t\}. \tag{5.18}$$

(b) If in (a) we take $\nu = \pi$, where π is the stationary distribution, then (by definition) $\pi P^t = \pi$, and (5.18) bounds the difference between μP^t and π. The only thing left to check is that there exists a coupling guaranteed to coalesce, that is, for which $\mathbf{P}\{\tau_{\text{couple}} < \infty\} = 1$. Show that if the chains (X_t) and (Y_t) are taken to be independent of one another, then they are assured to eventually meet.

EXERCISE 5.2. Let (X_t, Y_t) be a Markovian coupling such that for some $0 < \alpha < 1$ and some $t_0 > 0$, the coupling time $\tau_{\text{couple}} = \min\{t \geq 0 : X_t = Y_t\}$ satisfies $\mathbf{P}\{\tau_{\text{couple}} \leq t_0\} \geq \alpha$ for *all* pairs of initial states (x, y). Prove that

$$\mathbf{E}(\tau_{\text{couple}}) \leq \frac{t_0}{\alpha}.$$

EXERCISE 5.3. Show that if $X_1, X_2, \ldots$ are independent and each have mean μ and if τ is a $\mathbb{Z}^+$-valued random variable independent of all the X_i's, then

$$\mathbf{E}\left(\sum_{i=1}^{\tau} X_i\right) = \mu \mathbf{E}(\tau).$$

EXERCISE 5.4. We can get a better bound on the mixing time for the lazy walker on the d-dimensional torus by sharpening the analysis of the "coordinate-by-coordinate" coupling given in the proof of Theorem 5.5.

Let $t \geq kdn^2$.

(a) Show that the probability that the first coordinates of the two walks have not yet coupled by time t is less than $(1/4)^k$.
(b) By making an appropriate choice of k and considering all the coordinates, obtain an $O((d \log d)n^2)$ bound on t_{mix}.

Notes

The use of coupling in probability is usually traced back to Doeblin (1938). Couplings of Markov chains were first studied in Pitman (1974) and Griffeath (1974/75). See also Pitman (1976). See Luby, Randall, and Sinclair (1995) and Luby, Randall, and Sinclair (2001) for interesting examples of couplings.

For Glauber dynamics on colorings, it is shown in Chapter 14 that if the number of colors q satisfies $q > 2\Delta$, then the mixing time is of order $n \log n$.

Luby and Vigoda (1999) show that for a different Markov chain with the hardcore model as its stationary distribution, for λ small enough, the mixing time is of order $n \log n$. See also Luby and Vigoda (1995) and Vigoda (2001).

Further reading. For more on coupling and its applications in probability, see Lindvall (2002) and Thorisson (2000).

CHAPTER 6

Strong Stationary Times

6.1. Top-to-Random Shuffle

We begin this chapter with an example. Consider the following (slow) method of shuffling a deck of n cards: take the top card and insert it uniformly at random in the deck. This process will eventually mix up the deck—the successive arrangements of the deck are a random walk on the group $\mathcal{S}_n$ of $n!$ possible permutations of the cards, which by Proposition 2.12 has uniform stationary distribution.

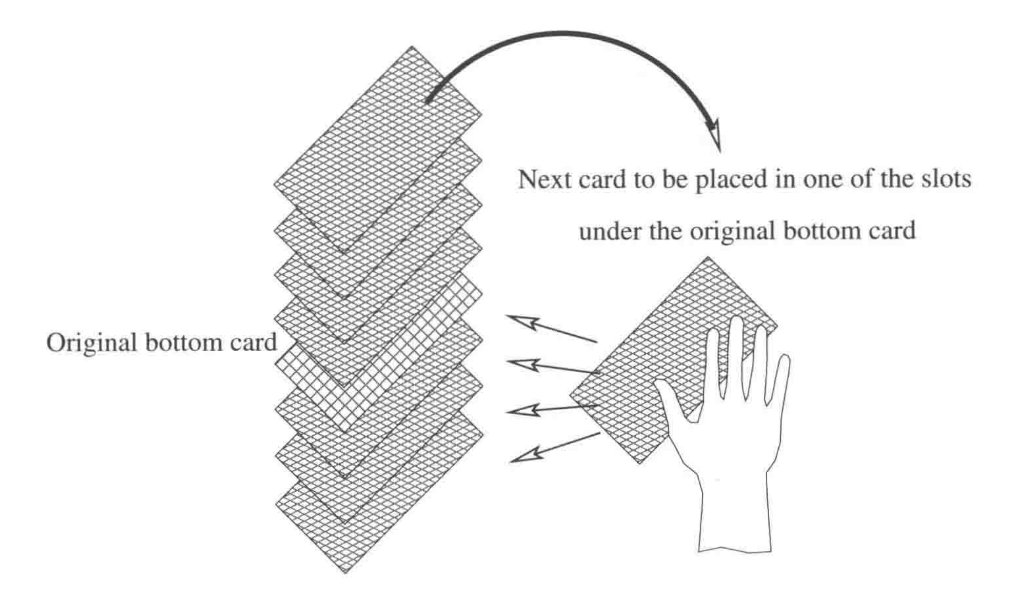


FIGURE 6.1. The top-to-random shuffle.

How long must we shuffle using this method until the arrangement of the deck is close to random?

Let τ_{top} be the time *one move after the first occasion when the original bottom card has moved to the top of the deck.* We show now that the arrangement of cards at time τ_{top} is distributed uniformly on the set $\mathcal{S}_n$ of all permutations of $\{1, \dots, n\}$ and moreover this random element of $\mathcal{S}_n$ is independent of the time τ_{top}.

More generally, we prove the following:

PROPOSITION 6.1. *Let (X_t) be the random walk on $\mathcal{S}_n$ corresponding to the top-to-random shuffle on n cards. Given at time t that there are k cards under the original bottom card, each of the k! possible orderings of these cards are equally likely. Therefore, if τ_{top} is one shuffle after the first time that the original bottom*

card moves to the top of the deck, then the distribution of $X_{\tau_{\text{top}}}$ is uniform over $\mathcal{S}_n$, and the time τ_{top} is independent of $X_{\tau_{\text{top}}}$.

PROOF. When $t = 0$, there are no cards under the original bottom card, and the claim is trivially valid. Now suppose that the claim holds at time t. There are two possibilities at time $t + 1$: either a card is placed under the original bottom card, or not. In the second case, the cards under the original bottom card remain in random order. In the first case, given that the card is placed under the original bottom card, each of the $k + 1$ possible locations for the card is equally likely, and so each of the $(k + 1)!$ orderings are equiprobable. ■

If we stop shuffling precisely one shuffle after the original bottom card rises to the top of the deck for the first time, then the order of the cards at this time is exactly uniform over all possible arrangements. That is, $X_{\tau_{\text{top}}}$ has *exactly* the stationary distribution of the chain. In this chapter, we show how we can use the distribution of the *random* time τ_{top} to bound t_{mix}, the *fixed* number of steps needed for the distribution of the chain to be *approximately* stationary.

6.2. Definitions

6.2.1. Stopping times. Suppose you give instructions to your stock broker to sell a particular security when its value next drops below 32 dollars per share. This directive can be implemented by a computer program: at each unit of time, the value of the security is checked; if the value at that time is at least 32, no action is taken, while if the value is less than 32, the asset is sold and the program quits.

You would like to tell your broker to sell a stock at the first time its value equals its maximum value over its lifetime. However, this is not a reasonable instruction, because to determine on Wednesday whether or not to sell, the broker needs to know that on Thursday the value will not rise and in fact for the entire infinite future that the value will never exceed its present value. To determine the correct decision on Wednesday, the broker must be able to see into the future!

The first instruction is an example of a *stopping time*, which we will now define, while the second rule is not.

Given a sequence $(X_t)_{t=0}^{\infty}$ of Ω-valued random variables, a $\{0, 1, 2, \ldots, \infty\}$-valued random variable τ is a ***stopping time*** for (X_t) if, for each $t \in \{0, 1, \ldots\}$, there is a set $B_t \subset \Omega^{t+1}$ such that

$$\{\tau = t\} = \{(X_0, X_1, \ldots, X_t) \in B_t\}.$$

In other words, a random time τ is a stopping time if and only if the indicator function $\mathbf{1}_{\{\tau=t\}}$ is a function of the vector $(X_0, X_1, \ldots, X_t)$.

EXAMPLE 6.2 (Hitting times). Fix $A \subseteq \Omega$. The vector $(X_0, X_1, \ldots, X_t)$ determines whether a site in A is visited for the first time at time t. That is, if

$$\tau_A = \min\{t \geq 0 \,:\, X_t \in A\}$$

is the first time that the sequence (X_t) is in A, then

$$\{\tau_A = t\} = \{X_0 \notin A, X_1 \notin A, \ldots, X_{t-1} \notin A, X_t \in A\}.$$

Therefore, τ_A is a stopping time. (We saw the special case where $A = \{x\}$ consists of a single state in Section 1.5.2.)

Consider the top-to-random shuffle, defined in Section 6.1. Let A be the set of arrangements having the original bottom card on top. Then $\tau_{\text{top}} = \tau_A + 1$. By Exercise 6.1, τ_{top} is a stopping time.

6.2.2. Randomized stopping times. The following example is instructive.

EXAMPLE 6.3 (Random walk on the hypercube). The lazy random walk $(\boldsymbol{X}_t)$ on the hypercube $\{0,1\}^n$ was introduced in Section 2.3, and we used coupling to bound the mixing time in Section 5.3.3. Recall that a move of this walk can be constructed using the following random mapping representation: an element (j, B) from $\{1, 2, \ldots, n\} \times \{0, 1\}$ is selected uniformly at random, and coordinate j of the current state is updated with the bit B.

In this construction, the chain is determined by the i.i.d. sequence (Z_t), where $Z_t = (j_t, B_t)$ is the coordinate and bit pair used to update at step t.

Define

$$\tau_{\text{refresh}} := \min\left\{t \geq 0 \,:\, \{j_1, \ldots, j_t\} = \{1, 2, \ldots, n\}\right\},$$

the first time when all the coordinates have been selected at least once for updating.

Because at time τ_{refresh} all of the coordinates have been replaced with independent fair bits, the distribution of the chain at this time is uniform on $\{0,1\}^n$. That is, $X_{\tau_{\text{refresh}}}$ is an exact sample from the stationary distribution π.

Note that τ_{refresh} is not a function of (X_t), but it is a function of (Z_t). In particular, while τ_{refresh} is not a stopping time for (X_t), it is a stopping time for (Z_t).

Recall that we showed in Section 1.2 that every transition matrix P has a random mapping representation: we can find an i.i.d. sequence $(Z_t)_{t=1}^{\infty}$ and a map f such that the sequence $(X_t)_{t=0}^{\infty}$ defined inductively by

$$X_0 = x, \qquad X_t = f(X_{t-1}, Z_t)$$

is a Markov chain with transition matrix P started from x. A random time τ is called a ***randomized stopping time*** for the Markov chain (X_t) if it is a stopping time for the sequence (Z_t).

EXAMPLE 6.4. We return to Example 6.3, the lazy random walk on the hypercube. As remarked there, the time τ_{refresh} is a stopping time for the sequence (Z_t), where Z_t is the coordinate and bit used to update at time t. Therefore, τ_{refresh} is a randomized stopping time.

6.3. Achieving Equilibrium

For the top-to-random shuffle, one shuffle after the original bottom card rises to the top, the deck is in completely random order. Likewise, for the lazy random walker on the hypercube, at the first time when all of the coordinates have been updated, the state of the chain is a random sample from $\{0,1\}^n$. These random times are examples of ***stationary times***, which we now define.

Let (X_t) be an irreducible Markov chain with stationary distribution π. A ***stationary time*** τ for (X_t) is a randomized stopping time, possibly depending on the starting position x, such that the distribution of X_τ is π:

$$\mathbf{P}_x\{X_\tau = y\} = \pi(y). \tag{6.1}$$

EXAMPLE 6.5. Let (X_t) be an irreducible Markov chain with state space Ω and stationary distribution π. Let ξ be a Ω-valued random variable with distribution π, and define

$$\tau = \min\{t \geq 0 : X_t = \xi\}.$$

The time τ is a randomized stopping time, and because $X_\tau = \xi$, it follows that τ is a stationary time.

Suppose the chain starts at x_0. If $\tau = 0$, then $X_\tau = x_0$; therefore, τ and X_τ are not independent.

EXAMPLE 6.6. Let (X_t) be the random walk on the n-cycle. Define τ by tossing a coin with probability of heads $1/n$. If "heads", let $\tau = 0$; if "tails", let τ be the first time every state has been visited at least once. Given "tails", the distribution of X_τ is uniform over all $n-1$ states different from the starting state. (See Exercise 6.9.) This shows that X_τ has the uniform distribution, whence τ is a stationary time.

However, $\tau = 0$ implies that X_τ is the starting state. Therefore, as in Example 6.5, τ and X_τ are not independent.

As mentioned at the end of Section 6.1, we want to use the time τ_{top} to bound t_{mix}. To carry out this program, we need a property of τ_{top} stronger than (6.1). We will need that τ_{top} is independent of $X_{\tau_{\text{top}}}$, a property not enjoyed by the stationary times in Example 6.5 and Example 6.6.

6.4. Strong Stationary Times and Bounding Distance

A ***strong stationary time*** for a Markov chain (X_t) with stationary distribution π is a randomized stopping time τ, possibly depending on the starting position x, such that

$$\mathbf{P}_x\{\tau = t,\, X_\tau = y\} = \mathbf{P}_x\{\tau = t\}\pi(y). \tag{6.2}$$

In words, X_τ has distribution π and is independent of τ.

EXAMPLE 6.7. For the top-to-random shuffle, the first time τ_{top} when the original bottom card gets placed into the deck by a shuffle is a strong stationary time. This is the content of Proposition 6.1.

EXAMPLE 6.8. We return to Example 6.3, the lazy random walk on the hypercube. The time τ_{refresh}, the first time each of the coordinates have been refreshed with an independent fair bit, is a strong stationary time.

We now return to the program suggested at the end of Section 6.1 and use strong stationary times to bound t_{mix}.

We first need the following technical lemma.

LEMMA 6.9. *Let (X_t) be an irreducible Markov chain with stationary distribution π. If τ is a strong stationary time for (X_t), then for all $t \geq 0$,*

$$\mathbf{P}_x\{\tau \leq t,\, X_t = y\} = \mathbf{P}\{\tau \leq t\}\pi(y). \tag{6.3}$$

PROOF. Let $Z_1, Z_2, \ldots$ be the i.i.d. sequence used in the random mapping representation of (X_t). For any $s \leq t$,

$$\mathbf{P}_x\{\tau = s,\, X_t = y\} = \sum_{z \in \Omega} \mathbf{P}_x\{X_t = y \mid \tau = s,\, X_s = z\}\mathbf{P}_x\{\tau = s,\, X_s = z\}. \tag{6.4}$$

Since τ is a stopping time for (Z_t), the event $\{\tau = s\}$ equals $\{(Z_1, \ldots, Z_s) \in B\}$ for some set $B \subset \Omega^s$. Also, for integers $r, s \geq 0$, there exists a function $\tilde{f}_r : \Omega^{r+1} \to \Omega$ such that

$$X_{s+r} = \tilde{f}_r(X_s, Z_{s+1}, \ldots, Z_{s+r}).$$

Since the random vectors $(Z_1, \ldots, Z_s)$ and $(Z_{s+1}, \ldots, Z_t)$ are independent,

$$\begin{aligned}\mathbf{P}_x\{X_t = y \mid \tau = s,\, X_s = z\} \\ = \mathbf{P}_x\{\tilde{f}_{t-s}(z, Z_{s+1}, \ldots, Z_t) = y \mid (X_1, \ldots, X_s) \in B,\, X_s = z\} = P^{t-s}(z, y).\end{aligned}$$

Therefore, using the definition (6.2) along with the above equality, (6.4) can be rewritten as

$$\mathbf{P}_x\{\tau = s, X_t = y\} = \sum_{z \in \Omega} P^{t-s}(z, y)\pi(z)\mathbf{P}_x\{\tau = s\}. \tag{6.5}$$

Since π satisfies $\pi = \pi P^{t-s}$, the right-hand side of (6.5) equals $\pi(y)\mathbf{P}_x\{\tau = s\}$. Summing over $s \leq t$ establishes (6.3). ∎

The route from strong stationary times to bounding convergence time is the following proposition:

PROPOSITION 6.10. *If τ is a strong stationary time, then*

$$d(t) = \max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \max_{x \in \Omega} \mathbf{P}_x\{\tau > t\}. \tag{6.6}$$

We break the proof into two lemmas. It will be convenient to introduce a parameter $s_x(t)$, called ***separation distance*** and defined by

$$s_x(t) := \max_{y \in \Omega} \left[1 - \frac{P^t(x, y)}{\pi(y)}\right]. \tag{6.7}$$

We also define

$$s(t) := \max_{x \in \Omega} s_x(t). \tag{6.8}$$

The relationship between $s_x(t)$ and strong stationary times is

LEMMA 6.11. *If τ is a strong stationary time, then*

$$s_x(t) \leq \mathbf{P}_x\{\tau > t\}. \tag{6.9}$$

PROOF. Fix $x \in \Omega$. Observe that for any $y \in \Omega$,

$$1 - \frac{P^t(x, y)}{\pi(y)} = 1 - \frac{\mathbf{P}_x\{X_t = y\}}{\pi(y)} \leq 1 - \frac{\mathbf{P}_x\{X_t = y,\, \tau \leq t\}}{\pi(y)}. \tag{6.10}$$

By Lemma 6.9, the right-hand side equals

$$1 - \frac{\pi(y)\mathbf{P}_x\{\tau \leq t\}}{\pi(y)} = \mathbf{P}_x\{\tau > t\}. \tag{6.11}$$

∎

REMARK 6.12. Given starting state x, a state y is a ***halting state*** for a stopping time τ if $X_t = y$ implies $\tau \leq t$. For example, when starting the lazy random walk on the hypercube at $(0, \ldots, 0)$, the state $(1, \ldots, 1)$ is a halting state for the stopping time τ_{refresh} defined in Example 6.3. Because the inequality in (6.10) is an equality if and only if y is a halting state for the starting state x, it follows that the inequality in (6.9) is an equality if and only if there exists a halting state for the starting state x.

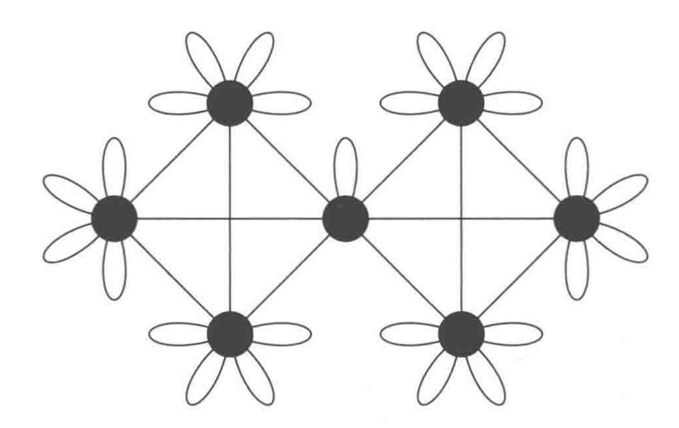


FIGURE 6.2. Two complete graphs (on 4 vertices), "glued" at a single vertex. Loops have been added so that every vertex has the same degree (count each loop as one edge).

The next lemma along with Lemma 6.11 proves Proposition 6.10.

LEMMA 6.13. *The separation distance* $s_x(t)$ *satisfies*

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq s_x(t), \tag{6.12}$$

and therefore $d(t) \leq s(t)$.

PROOF. We have

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} = \sum_{\substack{y\in\Omega \\ P^t(x,y)<\pi(y)}} \left[\pi(y) - P^t(x,y)\right] = \sum_{\substack{y\in\Omega \\ P^t(x,y)<\pi(y)}} \pi(y)\left[1 - \frac{P^t(x,y)}{\pi(y)}\right]$$
$$\leq \max_y \left[1 - \frac{P^t(x,y)}{\pi(y)}\right] = s_x(t).$$

■

6.5. Examples

6.5.1. Two glued complete graphs. Consider the graph G obtained by taking two complete graphs on n vertices and "gluing" them together at a single vertex. We analyze here simple random walk on a slightly modified graph, G'.

Let $v^\star$ be the vertex where the two complete graphs meet. After gluing, $v^\star$ has degree $2n-2$, while every other vertex has degree $n-1$. To make the graph regular and to ensure non-zero holding probability at each vertex, in G' we add one loop at $v^\star$ and n loops at all other vertices. (See Figure 6.2 for an illustration when $n = 4$.) The uniform distribution is stationary for simple random walk on G', since it is regular of degree $2n-1$.

It is clear that when at $v^\star$, the next state is equally likely to be any of the $2n-1$ vertices. For this reason, if τ is the time one step after $v^\star$ has been visited for the first time, then τ is a strong stationary time.

When the walk is *not* at $v^\star$, the probability of moving (in one step) to $v^\star$ is $1/(2n-1)$. This remains true at any subsequent move. That is, the first time $\tau_{v^\star}$

that the walk visits $v^\star$ is geometric with $\mathbf{E}(\tau_{v^\star}) = 2n - 1$. Therefore, $\mathbf{E}(\tau) = 2n$, and using Markov's inequality shows that

$$\mathbf{P}_x\{\tau \geq t\} \leq \frac{2n}{t}. \tag{6.13}$$

Taking $t = 8n$ in (6.13) and applying Proposition 6.10 shows that

$$t_{\text{mix}} \leq 8n.$$

A lower bound on t_{mix} of order n is obtained in Exercise 6.7.

6.5.2. Random walk on the hypercube. We return to Example 6.3, the lazy random walker on $\{0,1\}^n$. As noted in Example 6.8, the random variable τ_{refresh}, the time when each coordinate has been selected at least once for the first time, is a strong stationary time. The time τ_{refresh} and the coupling time τ_{couple} for the coordinate-by-coordinate coupling used in Section 5.3.3 are closely related: the coupon collector's time of Section 2.2 stochastically dominates τ_{couple} and has the same distribution as τ_{refresh}. It is therefore not surprising that we obtain here exactly the same upper bound for t_{mix} as was found using the coupling method. In particular, combining Proposition 2.4 and Lemma 6.11 shows that the separation distance satisfies, for each x,

$$s_x(n \log n + cn) \leq e^{-c}. \tag{6.14}$$

By Lemma 6.13,

$$t_{\text{mix}}(\varepsilon) \leq n \log n + \log(\varepsilon^{-1})n. \tag{6.15}$$

REMARK 6.14. The reason we explicitly give a bound on the separation distance here and appeal to Lemma 6.13, instead of applying directly Proposition 6.10, is that there is a matching lower bound on $s(t)$, which we give in Section 18.4. This contrasts with the lower bound on $d(t)$ we will find in Section 7.3.1, which implies $t_{\text{mix}}(1-\varepsilon) \geq (1/2)n \log n - c(\varepsilon)n$. In fact, the estimate on $t_{\text{mix}}(\varepsilon)$ given in (6.15) is off by a factor of two, as we will see in Section 18.2.2.

6.5.3. Top-to-random shuffle. We revisit the top-to-random shuffle introduced in Section 6.1. As noted in Example 6.7, the time τ_{top} is a strong stationary time.

Consider the motion of the original bottom card. When there are k cards beneath it, the chance that it rises one card remains k/n until a shuffle puts the top card underneath it. Thus, the distribution of τ_{top} is the same as the coupon collector's time. As above for the lazy hypercube walker, combining Proposition 6.10 and Proposition 2.4 yields

$$d(n \log n + \alpha n) \leq e^{-\alpha} \quad \text{for all } n. \tag{6.16}$$

Consequently,

$$t_{\text{mix}}(\varepsilon) \leq n \log n + \log(\varepsilon^{-1})n. \tag{6.17}$$

6.5.4. The move-to-front chain. A certain professor owns many books, arranged on his shelves. When he finishes with a book drawn from his collection, he does not waste time re-shelving it in its proper location. Instead, he puts it at the very beginning of his collection, in front of all the shelved books.

If his choice of book is random, this is an example of the ***move-to-front*** chain. It is a very natural chain which arises in many applied contexts. Any setting where



Drawing by Yelena Shvets

FIGURE 6.3. The move-to-front rule in action.

items are stored in a stack, removed at random locations, and placed on the top of the stack can be modeled by the move-to-front chain.

Let P be the transition matrix (on permutations of $\{1, 2, \ldots, n\}$) corresponding to this method of rearranging elements.

The time reversal $\widehat{P}$ of the move-to-front chain is the top-to-random shuffle, as intuition would expect. It is clear from the definition that for any permissible transition $\sigma_1 \to \sigma_2$ for move-to-front, the transition $\sigma_2 \to \sigma_1$ is permissible for top-to-random, and both have probability n^{-1}.

By Lemma 4.13, the mixing time for move-to-front will be identical to that of the top-to-random shuffle. Consequently, the mixing time for move-to-front is not more than $n \log n - \log(\varepsilon)n$.

6.5.5. Lazy random walk on cycle. Here is a recursive description of a strong stationary time τ_k for lazy random walk (X_t) on a cycle $\mathbb{Z}_n$ with $n = 2^k$ points.

For $k = 1$, waiting one step will do: $\tau_1 = 1$ with mean $m_1 = 1$. Suppose we have constructed τ_k already and are now given a cycle with 2^{k+1} points. Set $T_0 = 0$ and define $T_1 = t_1$ as the time it takes the lazy walk to make two ± 1 steps.

Then T_1 is a sum of two geometric(1/2) random variables and thus has mean 4. Given $T_1, \ldots, T_j$, define t_{j+1} as the time it takes the lazy random walk to make two steps of ± 1 after time T_j and let $T_{j+1} = T_j + t_{j+1}$. Observe that the process (X_{T_j}) for $j \geq 0$ is lazy random walk on the even points of the cycle. Therefore at time T_{τ_k} the location of $X_{T_{\tau_k}}$ is uniform among the even points on the 2^{k+1}-cycle, even if we condition on the value of T_{τ_k}. It follows that $\tau_{k+1} = T_{\tau_k} + 1$ is a strong stationary time for the lazy random walk on the 2^{k+1}-cycle. Exercise 6.8 completes the discussion by showing that $m_k = (4^k - 1)/3$, where $m_k = \mathbf{E}\tau_k$.

6.6. Stationary Times and Cesaro Mixing Time*

We have seen that *strong* stationary times fit naturally with separation distance and can be used to bound the mixing time. We now see that stationary times fit naturally with an alternative definition of mixing time.

Consider a finite chain (X_t) with transition matrix P and stationary distribution π on Ω. Given $t \geq 1$, suppose that we choose uniformly a time $\sigma \in \{0, 1, \ldots, t-1\}$, and run the Markov chain for σ steps. Then the state X_σ has distribution

$$\nu_x^t := \frac{1}{t} \sum_{s=0}^{t-1} P^s(x, \cdot). \tag{6.18}$$

The ***Cesaro mixing time*** $t_{\mathrm{mix}}^{\star}(\varepsilon)$ is defined as the first t such that for all $x \in \Omega$,

$$\|\nu_x^t - \pi\|_{\mathrm{TV}} \leq \varepsilon\,.$$

See Exercises 10.12 through 10.14 for some properties of the Cesaro mixing time.

The following general result due to Lovász and Winkler (1998) shows that the expectation of any stationary time yields an upper bound for $t_{\mathrm{mix}}^{\star}(1/4)$. Remarkably, this does not need reversibility or laziness. Lovász and Winkler also prove a converse of this result.

THEOREM 6.15. *Consider a finite chain with transition matrix P and stationary distribution π on Ω. If τ is a stationary time for the chain, then $t_{\mathrm{mix}}^{\star}(1/4) \leq 4\max_{x\in\Omega} \mathbf{E}_x(\tau) + 1$.*

PROOF. Denote by ν_x^t the Cesaro average (6.18). Since τ is a stationary time, so is $\tau + s$ for every $s \geq 1$. Therefore, for every $y \in \Omega$,

$$t\pi(y) = \sum_{s=0}^{t-1} \mathbf{P}_x\{X_{\tau+s} = y\} = \sum_{\ell=0}^{\infty} \mathbf{P}_x\{X_\ell = y,\, \tau \leq \ell < \tau + t\}\,.$$

Consequently,

$$t\nu_x^t(y) - t\pi(y) \leq \sum_{\ell=0}^{t-1} \mathbf{P}_x\{X_\ell = y,\, \tau > \ell\}\,.$$

Summing the last inequality over all $y \in \Omega$ such that the right-hand side is positive,

$$t\|\nu_x^t - \pi\|_{\mathrm{TV}} \leq \sum_{\ell=0}^{t-1} \mathbf{P}_x\{\tau > \ell\} \leq \mathbf{E}_x(\tau)\,.$$

Thus for $t \geq 4\mathbf{E}_x(\tau)$ we have $\|\nu_x^t - \pi\|_{\mathrm{TV}} \leq 1/4$. ∎

Exercises

EXERCISE 6.1. Show that if τ and τ' are stopping times for the sequence (X_t), then $\tau + \tau'$ is a stopping time for (X_t). In particular, if r is a non-random and non-negative integer and τ is a stopping time, then $\tau + r$ is a stopping time.

EXERCISE 6.2. Consider the top-to-random shuffle. Show that the time until the card initially one card from the bottom rises to the top, plus one more move, is a strong stationary time, and find its expectation.

EXERCISE 6.3. Show that for the Markov chain on two complete graphs in Section 6.5.1, the stationary distribution is uniform on all $2n - 1$ vertices.

EXERCISE 6.4. Let $s(t)$ be defined as in (6.8).

(a) Show that there is a stochastic matrix Q so that $P^t(x, \cdot) = [1 - s(t)]\,\pi + s(t)Q^t(x, \cdot)$ and $\pi = \pi Q$.
(b) Using the representation in (a), show that

$$P^{t+u}(x, y) = [1 - s(t)s(u)]\,\pi(y) + s(t)s(u) \sum_{z \in \Omega} Q^t(x, z)Q^u(z, y). \tag{6.19}$$

(c) Using (6.19), establish that s is submultiplicative: $s(t + u) \leq s(t)s(u)$.

EXERCISE 6.5. Show that if $\max_{x \in \Omega} \mathbf{P}_x\{\tau > t_0\} \leq \varepsilon$, then $d(t) \leq \varepsilon^{t/t_0}$.

EXERCISE 6.6 (Wald's Identity). Let (Y_t) be a sequence of independent and identically distributed random variables such that $\mathbf{E}(|Y_t|) < \infty$.

(a) Show that if τ is a random time so that the event $\{\tau \geq t\}$ is independent of Y_t and $\mathbf{E}(\tau) < \infty$, then

$$\mathbf{E}\left(\sum_{t=1}^{\tau} Y_t\right) = \mathbf{E}(\tau)\mathbf{E}(Y_1). \tag{6.20}$$

Hint: Write $\sum_{t=1}^{\tau} Y_t = \sum_{t=1}^{\infty} Y_t \mathbf{1}_{\{\tau \geq t\}}$. First consider the case where $Y_t \geq 0$.
(b) Let τ be a stopping time for the sequence (Y_t). Show that $\{\tau \geq t\}$ is independent of Y_t, so (6.20) holds provided that $\mathbf{E}(\tau) < \infty$.

EXERCISE 6.7. Consider the Markov chain of Section 6.5.1 defined on two glued complete graphs. By considering the set $A \subset \Omega$ of all vertices in one of the two complete graphs, show that $t_{\text{mix}} \geq (n/2)\,[1 + o(1)]$.

EXERCISE 6.8. Let τ_k be the stopping time constructed in Section 6.5.5, and let $m_k = \mathbf{E}(\tau_k)$. Show that $m_{k+1} = 4m_k + 1$, so that $m_k = \sum_{i=0}^{k-1} 4^i = (4^k - 1)/3$.

EXERCISE 6.9. For a graph G, let W be the (random) vertex occupied at the first time the random walk has visited every vertex. That is, W is the last new vertex to be visited by the random walk. Prove the following remarkable fact: for random walk on an n-cycle, W is uniformly distributed over all vertices different from the starting vertex.

REMARK 6.16. Let W be the random vertex defined in Exercise 6.9. Lovász and Winkler (1993) demonstrate that cycles and complete graphs are the only graphs for which W is this close to uniformly distributed. More precisely, these families are the only ones for which W is equally likely to be any vertex other than the starting state.

Notes

Strong stationary times were introduced in Aldous and Diaconis (1987); see also Aldous and Diaconis (1986). An important class of strong stationary times was constructed by Diaconis and Fill (1990). The thesis of Pak (1997) has many examples of strong stationary times.

Aldous and Diaconis (1987) showed that for reversible chains, the distances s and $\bar{d}$ are also related by

$$s(2t) \leq 1 - (1 - \bar{d}(t))^2. \tag{6.21}$$

(See Aldous and Fill (1999, Chapter 4, Lemma 7).) We prove this as Lemma 19.3.

Lovász and Winkler (1995b, Theorem 5.1) showed that a stationary time has minimal expectation among all stationary times if and only if it has a halting state. (See also Lovász and Winkler (1998).)

For the lazy random walk on the hypercube, the strong stationary time τ_{refresh} achieved the bound (6.9). Aldous and Diaconis (1987) prove that, for any irreducible finite Markov chain, given a state x, there always exists a strong stationary time τ such that $s(t) = \mathbf{P}_x\{\tau > t\}$ for all t.

The strong stationary time we give for the cycle in Section 6.5.5 is due to Diaconis and Fill (1990), although the exposition is different. The idea goes back to Dubins's construction of the Skorokhod embedding (Dubins, 1968).

CHAPTER 7

Lower Bounds on Mixing Times

To this point, we have directed our attention to finding upper bounds on t_{mix}. Rigorous upper bounds lend confidence that simulation studies or randomized algorithms perform as advertised. It is natural to ask if a given upper bound is the best possible, and so in this chapter we turn to methods of obtaining lower bounds on t_{mix}.

7.1. Counting and Diameter Bounds

7.1.1. Counting bound. If the possible locations of a chain after t steps do not form a significant fraction of the state space, then the distribution of the chain at time t cannot be close to uniform. This idea can be used to obtain lower bounds on the mixing time.

Let (X_t) be a Markov chain with irreducible and aperiodic transition matrix P on the state space Ω, and suppose that the stationary distribution π is uniform over Ω. Define $d_{\text{out}}(x) := |\{y : P(x,y) > 0\}|$ to be the number of states accessible in one step from x, and let

$$\Delta := \max_{x \in \Omega} d_{\text{out}}(x). \tag{7.1}$$

Denote by Ω_t^x the set of states accessible from x in t steps, and observe that $|\Omega_t^x| \leq \Delta^t$. If $\Delta^t < (1-\varepsilon)|\Omega|$, then from the definition of total variation distance we have that

$$\left\|P^t(x,\cdot) - \pi\right\|_{TV} \geq P_t(x, \Omega_t^x) - \pi(\Omega_t^x) \geq 1 - \frac{\Delta^t}{|\Omega|} > \varepsilon.$$

This implies that

$$t_{\text{mix}}(\varepsilon) \geq \frac{\log(|\Omega|(1-\varepsilon))}{\log \Delta}. \tag{7.2}$$

Example 7.1 (Random walk on a d-regular graph). For random walk on a d-regular graph, the stationary distribution is uniform, so the inequality (7.2) can be applied. In this case, it yields the lower bound $t_{\text{mix}}(\varepsilon) \geq \log(|\Omega|(1-\varepsilon))/\log d$.

We use the bound (7.2) to bound below the mixing time for the riffle shuffle in Proposition 8.14.

7.1.2. Diameter bound. Given a transition matrix P on Ω, construct a graph with vertex set Ω and which includes the edge $\{x, y\}$ for all x and y with $P(x,y) + P(y,x) > 0$. Define the ***diameter*** of a Markov chain to be the diameter of this graph, that is, the maximal graph distance between distinct vertices.

Let P be an irreducible and aperiodic transition matrix on Ω with diameter L, and suppose that x_0 and y_0 are states at maximal graph distance L. Then

$P^{\lfloor (L-1)/2 \rfloor}(x_0, \cdot)$ and $P^{\lfloor (L-1)/2 \rfloor}(y_0, \cdot)$ are positive on disjoint vertex sets. Consequently, $\bar{d}(\lfloor (L-1)/2 \rfloor) = 1$ and for any $\varepsilon < 1/2$,

$$t_{\mathrm{mix}}(\varepsilon) \geq \frac{L}{2}. \tag{7.3}$$

7.2. Bottleneck Ratio

Bottlenecks in the state space Ω of a Markov chain are geometric features that control mixing time. A bottleneck makes portions of Ω difficult to reach from some starting locations, limiting the speed of convergence. Figure 7.1 is a sketch of a graph with an obvious bottleneck.

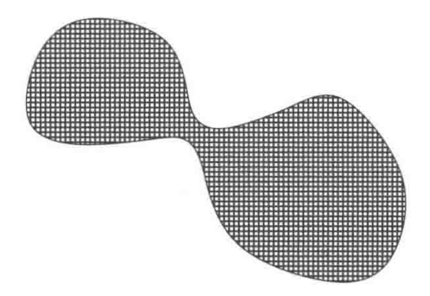


FIGURE 7.1. A graph with a bottleneck.

As usual, P is the irreducible and aperiodic transition matrix for a Markov chain on Ω with stationary distribution π.

The ***edge measure*** Q is defined by

$$Q(x, y) := \pi(x)P(x, y), \quad Q(A, B) = \sum_{x \in A, y \in B} Q(x, y). \tag{7.4}$$

Here $Q(A, B)$ is the probability of moving from A to B in one step when starting from the stationary distribution.

The ***bottleneck ratio*** of the set S is defined to be

$$\Phi(S) := \frac{Q(S, S^c)}{\pi(S)}, \tag{7.5}$$

while the bottleneck ratio of the whole chain is

$$\Phi_\star := \min_{S \,:\, \pi(S) \leq \frac{1}{2}} \Phi(S). \tag{7.6}$$

For simple random walk on a graph with vertices Ω and edge set E,

$$Q(x, y) = \begin{cases} \frac{\deg(x)}{2|E|} \frac{1}{\deg(x)} = \frac{1}{2|E|} & \text{if } \{x, y\} \text{ is an edge,} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $2|E|Q(S, S^c)$ is the size of the ***boundary*** ∂S of S, the collection of edges having one vertex in S and one vertex in S^c. The bottleneck ratio, in this case, becomes

$$\Phi(S) = \frac{|\partial S|}{\sum_{x \in S} \deg(x)}. \tag{7.7}$$

REMARK 7.2. If the walk is lazy, then $Q(x, y) = (4|E|)^{-1} \mathbf{1}_{\{\{x,y\} \in E\}}$, and the bottleneck ratio equals $\Phi(S) = 2|\partial S| / (\sum_{x \in S} \deg(x))$.

If the graph is regular with degree d, then $\Phi(S) = d^{-1}|\partial S|/|S|$, which is proportional to the ratio of the size of the boundary of S to the volume of S.

The relationship of $\Phi_\star$ to t_{mix} is the following theorem:

THEOREM 7.3. *If $\Phi_\star$ is the bottleneck ratio defined in* (7.6), *then*

$$t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4) \geq \frac{1}{4\Phi_\star}. \tag{7.8}$$

PROOF. Denote by π_S the restriction of π to S, so that $\pi_S(A) = \pi(A \cap S)$, and define μ_S to be π conditioned on S:

$$\mu_S(A) = \frac{\pi_S(A)}{\pi(S)}.$$

From Remark 4.3,

$$\pi(S)\left\|\mu_S P - \mu_S\right\|_{\mathrm{TV}} = \pi(S) \sum_{\substack{y\in\Omega,\\ \mu_S P(y)\geq \mu_S(y)}} \left[\mu_S P(y) - \mu_S(y)\right]. \tag{7.9}$$

Because $\pi_S P(y) = \pi(S)\mu_S P(y)$ and $\pi_S(y) = \pi(S)\mu_S(y)$, the inequality $\mu_S P(y) \geq \mu_S(y)$ holds if and only if $\pi_S P(y) \geq \pi_S(y)$. Thus

$$\pi(S)\left\|\mu_S P - \mu_S\right\|_{\mathrm{TV}} = \sum_{\substack{y\in\Omega,\\ \pi_S P(y)\geq \pi_S(y)}} \left[\pi_S P(y) - \pi_S(y)\right]. \tag{7.10}$$

Because $\pi_S(x) > 0$ only for $x \in S$ and $\pi_S(x) = \pi(x)$ for $x \in S$,

$$\pi_S P(y) = \sum_{x\in\Omega} \pi_S(x)P(x,y) = \sum_{x\in S} \pi(x)P(x,y) \leq \sum_{x\in\Omega} \pi(x)P(x,y) = \pi(y). \tag{7.11}$$

Again using that $\pi(y) = \pi_S(y)$ for $y \in S$, from (7.11) follows the inequality

$$\pi_S P(y) \leq \pi_S(y) \quad \text{for } y \in S. \tag{7.12}$$

On the other hand, because π_S vanishes on S^c,

$$\pi_S P(y) \geq 0 = \pi_S(y) \quad \text{for } y \in S^c. \tag{7.13}$$

Combining (7.12) and (7.13) shows that the sum on the right in (7.10) can be taken over S^c:

$$\pi(S)\left\|\mu_S P - \mu_S\right\|_{\mathrm{TV}} = \sum_{y\in S^c} \left[\pi_S P(y) - \pi_S(y)\right]. \tag{7.14}$$

Again because $\pi_S(y) = 0$ for $y \in S^c$,

$$\pi(S)\left\|\mu_S P - \mu_S\right\|_{\mathrm{TV}} = \sum_{y\in S^c}\sum_{x\in S} \pi(x)P(x,y) = Q(S,S^c).$$

Dividing by $\pi(S)$,

$$\left\|\mu_S P - \mu_S\right\|_{\mathrm{TV}} = \Phi(S).$$

By Exercise 4.3, for any $u \geq 0$,

$$\left\|\mu_S P^{u+1} - \mu_S P^u\right\|_{\mathrm{TV}} \leq \left\|\mu_S P - \mu_S\right\|_{\mathrm{TV}} = \Phi(S).$$

Using the triangle inequality on $\mu_S P^t - \mu_S = \sum_{u=0}^{t-1}(\mu_S P^{u+1} - \mu_S P^u)$ shows that

$$\left\|\mu_S P^t - \mu_S\right\|_{\mathrm{TV}} \leq t\Phi(S). \tag{7.15}$$

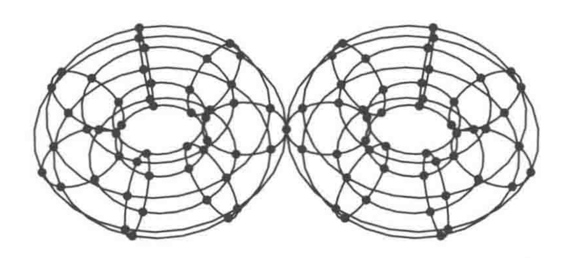


FIGURE 7.2. Two "glued" two-dimensional tori.

Assume that $\pi(S) \leq \frac{1}{2}$. In this case, because $\mu_S(S^c) = 0$,

$$\|\mu_S - \pi\|_{\mathrm{TV}} \geq \pi(S^c) - \mu_S(S^c) = 1 - \pi(S) \geq \frac{1}{2}.$$

Using the triangle inequality again shows that

$$\frac{1}{2} \leq \|\mu_S - \pi\|_{\mathrm{TV}} \leq \left\|\mu_S - \mu_S P^t\right\|_{\mathrm{TV}} + \left\|\mu_s P^t - \pi\right\|_{\mathrm{TV}}. \tag{7.16}$$

Taking $t = t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$ in (7.16), by the definition of t_{mix} and the inequality in (7.15),

$$\frac{1}{2} \leq t_{\mathrm{mix}}\Phi(S) + \frac{1}{4}.$$

Rearranging and minimizing over S establishes (7.8). ■

EXAMPLE 7.4 (Two glued tori). Consider the graph consisting of two d-dimensional tori "glued" together at a single vertex $v^\star$; see Figure 7.2 for an example of dimension two. Denote by V_1 and V_2 the sets of vertices in the right and left tori, respectively. Note that $V_1 \cap V_2 = v^\star$.

The set ∂V_1 consists of all edges $\{v^\star, v\}$, where $v \in V_2$. The size of ∂V_1 is $2d$. Also, $\sum_{x \in V_1} \deg(x) = 2dn^2 + 2d$. Consequently, the lazy random walk on this graph has

$$\Phi_\star \leq \Phi(V_1) = \frac{2(2d)}{2d(n^2+1)} \leq 2n^{-2}.$$

(See Remark 7.2.) Theorem 7.3 implies that $t_{\mathrm{mix}} \geq n^2/8$. We return to this example in Section 10.6, where it is proved that t_{mix} is of order $n^2 \log n$. Thus the lower bound here does not give the correct order.

EXAMPLE 7.5 (Coloring the star). Let Ω be the set of all proper q-colorings of a graph G, and let π be the uniform distribution on Ω. Recall from Example 3.5 that Glauber dynamics for π is the Markov chain which makes transitions as follows: at each unit of time, a vertex is chosen from V uniformly at random, and the color at this vertex is chosen uniformly at random from all ***feasible colors***. The feasible colors at vertex v are all colors *not* present among the neighbors of v.

We will prove (Theorem 14.8) that if $q > 2\Delta$, where Δ is the maximum degree of the graph, then the Glauber dynamics has mixing time of the order $|V| \log |V|$.

We show, by example, that quite different behavior may occur if the maximal degree is not bounded.

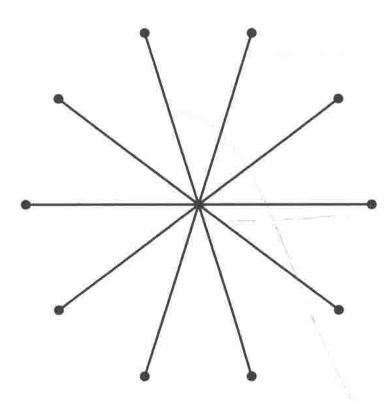


FIGURE 7.3. The star graph with 11 vertices.

The graph we study here is the ***star*** with n vertices, shown in Figure 7.3. This graph is a tree of depth 1 with $n-1$ leaves.

Let $v_\star$ denote the root vertex and let $S \subseteq \Omega$ be the set of proper colorings such that $v_\star$ has color 1:

$$S := \{x \in \Omega : x(v_\star) = 1\}.$$

For $(x, y) \in S \times S^c$, the edge measure $Q(x, y)$ is non-zero if and only if

- $x(v_\star) = 1$ and $y(v_\star) \neq 1$,
- $x(v) = y(v)$ for all leaves v, and
- $x(v) \notin \{1, y(v_\star)\}$ for all leaves v.

The number of such (x, y) pairs is therefore equal to $(q-1)(q-2)^{n-1}$, since there are $(q-1)$ possibilities for the color $y(v_\star)$ and $(q-2)$ possibilities for the color (identical in both x and y) of each of the $n-1$ leaves. Also, for such pairs, $Q(x, y) \leq (|\Omega| n)^{-1}$. It follows that

$$\sum_{x \in S, y \in S^c} Q(x, y) \leq \frac{1}{|\Omega| n}(q-1)(q-2)^{n-1}. \tag{7.17}$$

Since $x \in S$ if and only if $x(v_\star) = 1$ and $x(v) \neq 1$ for all $v \neq v_\star$, we have that $|S| = (q-1)^{n-1}$. Together with (7.17), this implies

$$\frac{Q(S, S^c)}{\pi(S)} = \frac{(q-1)(q-2)^{n-1}}{n(q-1)^{n-1}} = \frac{(q-1)^2}{n(q-2)}\left(1 - \frac{1}{q-1}\right)^n \leq \frac{(q-1)^2}{n(q-2)} e^{-n/(q-1)}.$$

Consequently, the mixing time is at least of exponential order:

$$t_{\text{mix}} \geq \frac{n(q-2)}{4(q-1)^2} e^{n/(q-1)}.$$

REMARK 7.6. In fact, this argument shows that if $n/(q \log q) \to \infty$, then t_{mix} is super-polynomial in n.

EXAMPLE 7.7 (Binary tree). Consider the lazy random walk on the rooted binary tree of depth k. (See Section 5.3.4 for the definition.) Let n be the number of vertices, so $n = 2^{k+1} - 1$. The number of edges is $n - 1$. In Section 5.3.4 we showed that $t_{\text{mix}} \leq 4n$. We now show that $t_{\text{mix}} \geq (n-2)/4$.

Let v_0 denote the root. Label the vertices adjacent to v_0 as v_r and v_ℓ. Call w a ***descendant*** of v if the shortest path from w to v_0 passes through v. Let S consist of the right-hand side of the tree, that is, v_r and all of its descendants.

We write $|v|$ for the length of the shortest path from v to v_0. By Example 1.12, the stationary distribution is

$$\pi(v) = \begin{cases} \frac{2}{2n-2} & \text{for } v = v_0, \\ \frac{3}{2n-2} & \text{for } 0 < |v| < k, \\ \frac{1}{2n-2} & \text{for } |v| = k. \end{cases}$$

Summing $\pi(v)$ over $v \in S$ shows that $\pi(S) = (n-2)/(2n-2)$. Since there is only one edge from S to S^c,

$$Q(S, S^c) = \pi(v_r)P(v_r, v_0) = \left(\frac{3}{2n-2}\right)\frac{1}{3} = \frac{1}{2n-2},$$

and so $\Phi(S) = 1/(n-2)$. Applying Theorem 7.3 establishes the lower bound

$$t_{\text{mix}} \geq \frac{n-2}{4} = \frac{2^{k+1}-3}{4},$$

which is exponentially large as a function of the depth k.

7.3. Distinguishing Statistics

One way to produce a lower bound on the mixing time t_{mix} is to find a statistic f (a real-valued function) on Ω such that the distance between the distribution of $f(X_t)$ and the distribution of f under the stationary distribution π can be bounded from below.

Let μ and ν be two probability distributions on Ω, and let f be a real-valued function defined on Ω. We write E_μ to indicate expectations of random variables (on sample space Ω) with respect to the probability distribution μ:

$$E_\mu(f) := \sum_{x \in \Omega} f(x)\mu(x).$$

(Note the distinction between E_μ with $\mathbf{E}_\mu$, the expectation operator corresponding to the Markov chain (X_t) started with initial distribution μ.) Likewise $\text{Var}_\mu(f)$ indicates variance computed with respect to the probability distribution μ.

PROPOSITION 7.8. *Let μ and ν be two probability distributions on Ω, and let f be a real-valued function on Ω. If*

$$|E_\mu(f) - E_\nu(f)| \geq r\sigma, \tag{7.18}$$

where $\sigma^2 = [\text{Var}_\mu(f) + \text{Var}_\nu(f)]/2$, then

$$\|\mu - \nu\|_{TV} \geq 1 - \frac{4}{4+r^2}. \tag{7.19}$$

Before proving this, we provide a useful lemma. When μ is a probability distribution on Ω and $f : \Omega \to \Lambda$, write μf^{-1} for the probability distribution defined by

$$(\mu f^{-1})(A) := \mu(f^{-1}(A))$$

for $A \subseteq \Lambda$. When X is an Ω-valued random variable with distribution μ, then $f(X)$ has distribution μf^{-1} on Λ.

LEMMA 7.9. *Let μ and ν be probability distributions on Ω, and let $f : \Omega \to \Lambda$ be a function on Ω, where Λ is a finite set. Then*

$$\|\mu - \nu\|_{TV} \geq \left\|\mu f^{-1} - \nu f^{-1}\right\|_{TV}.$$

PROOF. Since

$$|\mu f^{-1}(B) - \nu f^{-1}(B)| = |\mu(f^{-1}(B)) - \nu(f^{-1}(B))|,$$

it follows that

$$\max_{B \subset \Lambda} |\mu f^{-1}(B) - \nu f^{-1}(B)| \leq \max_{A \subset \Omega} |\mu(A) - \nu(A)|.$$

■

REMARK 7.10. Lemma 7.9 can be used to lower bound the distance of some chain from stationarity in terms of the corresponding distance for a projection (in the sense of Section 2.3.1) of that chain. To do so, take Λ to be the relevant partition of Ω.

If α is a probability distribution on a finite subset Λ of $\mathbb{R}$, the translation of α by c is the probability distribution α_c on $\Lambda + c$ defined by $\alpha_c(x) = \alpha(x - c)$. Total variation distance is ***translation invariant***: if α and β are two probability distributions on a finite subset Λ of $\mathbb{R}$, then $\|\alpha_c - \beta_c\|_{\mathrm{TV}} = \|\alpha - \beta\|_{\mathrm{TV}}$.

PROOF OF PROPOSITION 7.8. Suppose that α and β are probability distributions on a finite subset Λ of $\mathbb{R}$. Let

$$m_\alpha := \sum_{x \in \Lambda} x\alpha(x), \quad m_\beta := \sum_{x \in \Lambda} x\beta(x)$$

be the mean of α and β, respectively, and assume that $m_\alpha > m_\beta$. Let $M = (m_\alpha - m_\beta)/2$. By translating, we can assume that $m_\alpha = M$ and $m_\beta = -M$. Let $\eta = (\alpha + \beta)/2$, and define

$$r(x) := \frac{\alpha(x)}{\eta(x)}, \quad s(x) := \frac{\beta(x)}{\eta(x)}.$$

By Cauchy-Schwarz,

$$4M^2 = \left[\sum_{x \in \Lambda} x[r(x) - s(x)]\eta(x)\right]^2 \leq \sum_{x \in \Lambda} x^2\eta(x) \sum_{x \in \Lambda} [r(x) - s(x)]^2\eta(x). \tag{7.20}$$

If $\alpha = \mu f^{-1}, \beta = \nu f^{-1}$, and $\Lambda = f(\Omega)$, then $m_{\mu f^{-1}} = E_\mu(f)$, and (7.18) implies that $4M^2 \geq r^2\sigma^2$. Note that

$$\sum_{x \in \Lambda} x^2\eta(x) = \frac{m_\alpha^2 + \mathrm{Var}(\alpha) + m_\beta^2 + \mathrm{Var}(\beta)}{2} = M^2 + \sigma^2. \tag{7.21}$$

Since

$$|r(x) - s(x)| = 2\frac{|\alpha(x) - \beta(x)|}{\alpha(x) + \beta(x)} \leq 2,$$

we have

$$\sum_{x \in \Lambda} [r(x) - s(x)]^2\eta(x) \leq 2\sum_{x \in \Lambda} |r(x) - s(x)|\eta(x) = 2\sum_{x \in \Lambda} |\alpha(x) - \beta(x)|. \tag{7.22}$$

Putting together (7.20), (7.21), and (7.22) shows that

$$M^2 \leq (M^2 + \sigma^2) \|\alpha - \beta\|_{\mathrm{TV}},$$

and rearranging shows that

$$\|\alpha - \beta\|_{\mathrm{TV}} \geq 1 - \frac{\sigma^2}{\sigma^2 + M^2}.$$

If $4M^2 \geq r^2\sigma^2$, then

$$\|\alpha - \beta\|_{\mathrm{TV}} \geq 1 - \frac{4}{4+r^2}. \tag{7.23}$$

Using (7.23) now shows that

$$\left\|\mu f^{-1} - \nu f^{-1}\right\|_{\mathrm{TV}} \geq 1 - \frac{4}{4+r^2}.$$

This together with Lemma 7.9 establishes (7.19). ∎

REMARK 7.11. Applying Chebyshev's inequality yields a similar lower bound. Suppose $E_\mu(f) \leq E_\nu(f)$, let $\sigma_\star^2 := \max\{\mathrm{Var}_\mu(f), \mathrm{Var}_\nu(f)\}$, and suppose that

$$E_\nu(f) - E_\mu(f) \geq r\sigma_\star.$$

If $A = (E_\mu(f) + r\sigma_\star/2, \infty)$, then Chebyshev's inequality yields that

$$\mu f^{-1}(A) \leq \frac{4}{r^2} \quad \text{and} \quad \nu f^{-1}(A) \geq 1 - \frac{4}{r^2},$$

whence

$$\left\|\mu f^{-1} - \nu f^{-1}\right\|_{\mathrm{TV}} \geq 1 - \frac{8}{r^2}.$$

Thus, in the case of equal variances, the bound (7.19) is better than the one obtained via Chebyshev.

7.3.1. Random walk on hypercube. We use Proposition 7.8 to bound below the mixing time for the random walk on the hypercube, studied in Section 6.5.2.

First we record a simple lemma concerning the coupon collector problem introduced in Section 2.2.

LEMMA 7.12. *Consider the coupon collecting problem with n distinct coupon types, and let $I_j(t)$ be the indicator of the event that the j-th coupon has* not *been collected by time t. Let $R_t = \sum_{j=1}^n I_j(t)$ be the number of coupon types not collected by time t. The random variables $I_j(t)$ are negatively correlated, and letting $p = \left(1 - \frac{1}{n}\right)^t$, we have for $t \geq 0$*

$$\mathbf{E}(R_t) = np, \tag{7.24}$$

$$\mathrm{Var}(R_t) \leq np(1-p) \leq \frac{n}{4}. \tag{7.25}$$

PROOF. Since $I_j(t) = 1$ if and only if the first t coupons are not of type j, it follows that

$$\mathbf{E}\left(I_j(t)\right) = \left(1 - \frac{1}{n}\right)^t = p \quad \text{and} \quad \mathrm{Var}(I_j(t)) = p(1-p).$$

Similarly, for $j \neq k$,

$$\mathbf{E}\left(I_j(t) I_k(t)\right) = \left(1 - \frac{2}{n}\right)^t,$$

whence

$$\mathrm{Cov}(I_j(t), I_k(t)) = \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{1}{n}\right)^{2t} \leq 0.$$

From this (7.24) and (7.25) follow. ∎

PROPOSITION 7.13. *For the lazy random walk on the n-dimensional hypercube,*

$$d\left(\frac{1}{2}n\log n - \alpha n\right) \geq 1 - 8e^{-2\alpha+1}. \tag{7.26}$$

PROOF. Let $\mathbf{1}$ denote the vector of ones $(1, 1, \ldots, 1)$, and let $W(\boldsymbol{x}) = \sum_{i=1}^n x^i$ be the Hamming weight of $\boldsymbol{x} = (x^1, \ldots, x^n) \in \{0,1\}^n$. We will apply Proposition 7.8 with $f = W$. The position of the walker at time t, started at $\mathbf{1}$, is denoted by $\boldsymbol{X}_t = (X_t^1, \ldots, X_t^n)$.

As π is uniform on $\{0,1\}^n$, the distribution of the random variable W under π is binomial with parameters n and $p = 1/2$. In particular,

$$E_\pi(W) = \frac{n}{2}, \quad \mathrm{Var}_\pi(W) = \frac{n}{4}.$$

Let R_t be the number of coordinates not updated at least once by time t. When starting from $\mathbf{1}$, the conditional distribution of $W(\boldsymbol{X}_t)$ given $R_t = r$ is the same as that of $r + B$, where B is a binomial random variable with parameters $n - r$ and $1/2$. Consequently,

$$\mathbf{E}_{\mathbf{1}}(W(\boldsymbol{X}_t) \mid R_t) = R_t + \frac{(n - R_t)}{2} = \frac{1}{2}(R_t + n).$$

By (7.24),

$$\mathbf{E}_{\mathbf{1}}(W(\boldsymbol{X}_t)) = \frac{n}{2}\left[1 + \left(1 - \frac{1}{n}\right)^t\right].$$

Using the identity $\mathrm{Var}(W(\boldsymbol{X}_t)) = \mathrm{Var}(\mathbf{E}(W(\boldsymbol{X}_t) \mid R_t)) + \mathbf{E}(\mathrm{Var}(W(\boldsymbol{X}_t) \mid R_t))$,

$$\mathrm{Var}_{\mathbf{1}}(W(\boldsymbol{X}_t)) = \frac{1}{4}\mathrm{Var}(R_t) + \frac{1}{4}[n - \mathbf{E}_{\mathbf{1}}(R_t)].$$

By Lemma 7.12, R_t is the sum of negatively correlated indicators and consequently $\mathrm{Var}(R_t) \leq \mathbf{E}(R_t)$. We conclude that

$$\mathrm{Var}_{\mathbf{1}}(W(\boldsymbol{X}_t)) \leq \frac{n}{4}.$$

Setting

$$\sigma = \sqrt{\max\{\mathrm{Var}_\pi(W), \mathrm{Var}_{\mathbf{1}}(W(\boldsymbol{X}_t))\}} = \frac{\sqrt{n}}{2},$$

we have

$$\begin{aligned}
|E_\pi(W) - \mathbf{E}_{\mathbf{1}}(W(\boldsymbol{X}_t))| &= \frac{n}{2}\left(1 - \frac{1}{n}\right)^t \\
&= \sigma\sqrt{n}\left(1 - \frac{1}{n}\right)^t \\
&= \sigma\exp\left\{-t[-\log(1 - n^{-1})] + \frac{\log n}{2}\right\} \\
&\geq \sigma\exp\left\{-\frac{t}{n}\left(1 + \frac{1}{n}\right) + \frac{\log n}{2}\right\}.
\end{aligned}$$

The inequality follows since $\log(1 - x) \geq -x - x^2$ for $0 \leq x \leq 1/2$. By Proposition 7.8,

$$\left\|P^t(\mathbf{1}, \cdot) - \pi\right\|_{\mathrm{TV}} \geq 1 - 8\exp\left\{\frac{2t}{n}\left(1 + \frac{1}{n}\right) - \log n\right\}. \tag{7.27}$$

The inequality (7.26) follows because

$$\frac{1}{2}n\log n - \alpha n \le t_n = \left[\frac{1}{2}n\log n - \left(\alpha - \frac{1}{2}\right)n\right]\left[1 - \frac{1}{n+1}\right],$$

and the right-hand side of (7.27) evaluated at $t = t_n$ is equal to $1 - 8e^{-2\alpha+1}$. ■

7.4. Examples

7.4.1. Random walk on the cycle. We return to the lazy random walk on the cycle (see Example 1.8 and Example 2.10). The upper bound $t_{\text{mix}} \le n^2$ was found in Section 5.3.2.

We complement this by giving a lower bound of the same order. We can couple (X_t) to (S_t), a lazy simple random walk on all of $\mathbb{Z}$, so that $X_t = S_t$ until τ, the first time that $|X_t|$ hits $n/2$. Then

$$\mathbf{P}\left\{\sup_{t\le\alpha n^2} |X_t| > n/4\right\} = \mathbf{P}\left\{\sup_{t\le\alpha n^2} |S_t| > n/4\right\} \le \mathbf{P}\left\{|S_{\alpha n^2}| > n/4\right\} \le c_1\alpha,$$

by Chebyshev's inequality. For $\alpha < \alpha_0$, where α_0 is small enough, the right-hand side is less than 1/8. If $A_n = \{k \in \mathbb{Z}_n : |k| \ge n/4\}$, then $\pi(A_n) \ge 1/2$, and

$$d(\alpha_0 n^2) \ge 1/2 - 1/8 > 1/4,$$

so $t_{\text{mix}} \ge \alpha_0 n^2$.

7.4.2. Top-to-random shuffle. The top-to-random shuffle was introduced in Section 6.1 and upper bounds on $d(t)$ and t_{mix} were obtained in Section 6.5.3. Here we obtain matching lower bounds.

The bound below, from Aldous and Diaconis (1986), uses only the definition of total variation distance.

PROPOSITION 7.14. *Let (X_t) be the top-to-random chain on n cards. For any $\varepsilon > 0$, there exists a constant α_0 such that $\alpha > \alpha_0$ implies that for all sufficiently large n,*

$$d_n(n\log n - \alpha n) \ge 1 - \varepsilon. \tag{7.28}$$

In particular, there is a constant α_1 such that for all sufficiently large n,

$$t_{\text{mix}} \ge n\log n - \alpha_1 n. \tag{7.29}$$

PROOF. The bound is based on the events

$$A_j = \{\text{the original bottom } j \text{ cards are in their original relative order}\}. \tag{7.30}$$

Let id be the identity permutation; we will bound $\|P^t(\text{id},\cdot) - \pi\|_{TV}$ from below.

Let τ_j be the time required for the card initially j-th from the bottom to reach the top. Then

$$\tau_j = \sum_{i=j}^{n-1} \tau_{j,i},$$

where $\tau_{j,i}$ is the time it takes the card initially j-th from the bottom to ascend from position i (from the bottom) to position $i+1$. The variables $\{\tau_{j,i}\}_{i=j}^{n-1}$ are

independent and $\tau_{j,i}$ has a geometric distribution with parameter $p = i/n$, whence $\mathbf{E}(\tau_{j,i}) = n/i$ and $\mathrm{Var}(\tau_{j,i}) < n^2/i^2$. We obtain the bounds

$$\mathbf{E}(\tau_j) = \sum_{i=j}^{n-1} \frac{n}{i} \geq n(\log n - \log j - 1) \tag{7.31}$$

and

$$\mathrm{Var}(\tau_j) \leq n^2 \sum_{i=j}^{\infty} \frac{1}{i(i-1)} \leq \frac{n^2}{j-1}. \tag{7.32}$$

Using the bounds (7.31) and (7.32), together with Chebyshev's inequality, yields

$$\begin{aligned}\mathbf{P}\{\tau_j < n\log n - \alpha n\} &\leq \mathbf{P}\{\tau_j - \mathbf{E}(\tau_j) < -n(\alpha - \log j - 1)\} \\ &\leq \frac{1}{(j-1)},\end{aligned}$$

provided that $\alpha \geq \log j + 2$. Define $t_n(\alpha) = n\log n - \alpha n$. If $\tau_j \geq t_n(\alpha)$, then the original j bottom cards remain in their original relative order at time $t_n(\alpha)$, so

$$P^{t_n(\alpha)}(\mathrm{id}, A_j) \geq \mathbf{P}\{\tau_j \geq t_n(\alpha)\} \geq 1 - \frac{1}{(j-1)},$$

for $\alpha \geq \log j + 2$. On the other hand, for the uniform stationary distribution

$$\pi(A_j) = 1/(j!) \leq (j-1)^{-1},$$

whence, for $\alpha \geq \log j + 2$,

$$d_n(\,t_n(\alpha)\,) \geq \left\| P^{t_n(\alpha)}(\mathrm{id}, \cdot) - \pi \right\|_{\mathrm{TV}} \geq P^{t_n(\alpha)}(\mathrm{id}, A_j) - \pi(A_j) > 1 - \frac{2}{j-1}. \tag{7.33}$$

Taking $j = e^{\alpha-2}$, provided $n \geq e^{\alpha-2}$, we have

$$d_n(\,t_n(\alpha)\,) > g(\alpha) := 1 - \frac{2}{e^{\alpha-2}-1}.$$

Therefore,

$$\liminf_{n\to\infty} d_n(\,t_n(\alpha)\,) \geq g(\alpha),$$

where $g(\alpha) \to 1$ as $\alpha \to \infty$. ■

7.4.3. East model. Let

$$\Omega := \{x \in \{0,1\}^{n+1} \,:\, x(n+1) = 1\}.$$

The ***East model*** is the Markov chain on Ω which moves from x by selecting a coordinate k from $\{1, 2, \ldots, n\}$ at random and flipping the value $x(k)$ at k if and only if $x(k+1) = 1$. The reader should check that the uniform measure on Ω is stationary for these dynamics.

THEOREM 7.15. *For the East model, $t_{\mathrm{mix}} \geq n^2 - 2n^{3/2}$.*

PROOF. If $A = \{x : x(1) = 1\}$, then $\pi(A) = 1/2$.

On the other hand, we now show that it takes order n^2 steps until $X_t(1) = 1$ with probability near $1/2$ when starting from $x_0 = (0, 0, \ldots, 0, 1)$. Consider the motion of the left-most 1: it moves to the left by one if and only if the site immediately to its left is chosen. Thus, the waiting time for the left-most 1 to move from k to $k-1$ is bounded below by a geometric random variable G_k with mean

n. The sum $G = \sum_{k=1}^{n} G_k$ has mean n^2 and variance $(1 - n^{-1})n^3$. Therefore, if $t(n, \alpha) = n^2 - \alpha n^{3/2}$, then

$$\mathbf{P}\{X_{t(n,\alpha)}(1) = 1\} \leq \mathbf{P}\{G - n^2 \leq -\alpha n^{3/2}\} \leq \frac{1}{\alpha^2},$$

and so

$$|P^{t(n,\alpha)}(x_0, A) - \pi| \geq \frac{1}{2} - \frac{1}{\alpha^2}.$$

Thus, if $t \leq n^2 - 2n^{3/2}$, then $d(t) \geq 1/4$. In other words, $t_{\text{mix}} \geq n^2 - 2n^{3/2}$. ■

Exercises

EXERCISE 7.1. Let $\boldsymbol{X}_t = (X_t^1, \ldots, X_t^n)$ be the position of the lazy random walker on the hypercube $\{0,1\}^n$, started at $\boldsymbol{X}_0 = \mathbf{1} = (1, \ldots, 1)$. Show that the covariance between X_t^i and X_t^j is negative. Conclude that if $W(\boldsymbol{X}_t) = \sum_{i=1}^{n} X_t^i$, then $\text{Var}(W(X_t)) \leq n/4$.

Hint: It may be easier to consider the variables $Y_t^i = 2X_t^i - 1$.

EXERCISE 7.2. Show that $Q(S, S^c) = Q(S^c, S)$ for any $S \subset \Omega$. (This is easy in the reversible case, but holds generally.)

EXERCISE 7.3. An ***empty graph*** has no edges. Show that there is a constant $c(q)$ so that Glauber dynamics on the set of proper colorings of the empty graph satisfies

$$t_{\text{mix}} \geq \frac{1}{2} n \log n - c(q)n.$$

Hint: Copy the idea of the proof of Proposition 7.13.

Notes

The bottleneck ratio $\Phi_\star$ has many names in the literature, including *conductance*, *Cheeger constant*, and *isoperimetric constant*. It is more common to relate $\Phi_\star$ to the *spectral gap* of a Markov chain. This connection is discussed in Chapter 12. The approach to the lower bound for t_{mix} presented here is more direct and avoids reversibility. Results related to Theorem 7.3 can be found in Mihail (1989), Fill (1991), and Chen, Lovász, and Pak (1999).

Hayes and Sinclair (2007) have recently shown that the Glauber dynamics for many stationary distributions, on graphs of bounded degree, have mixing time order $n \log n$.

Upper bounds on the relaxation time (see Section 12.2) for the East model are obtained in Aldous and Diaconis (2002), which imply that $t_{\text{mix}} = O(n^2)$. See also Cancrini, Martinelli, Roberto, and Toninelli (2008) for results concerning a class of models including the East model. For combinatorics related to the East model, see Chung, Diaconis, and Graham (2001).

CHAPTER 8

The Symmetric Group and Shuffling Cards

> ...to destroy all organization far more shuffling is necessary than one would naturally suppose; I learned this from experience during a period of addiction, and have since compared notes with others.
>
> —Littlewood (1948).

We introduced the top-to-random shuffle in Section 6.1 and gave upper and lower bounds on its mixing time in Sections 6.5.3 and Section 7.4.2, respectively. Here we describe a general mathematical model for shuffling mechanisms and study two natural methods of shuffling cards.

We will return in Chapter 16 to the subject of shuffling, armed with techniques developed in intervening chapters. While games of chance have motivated probabilists from the founding of the field, there are several other motivations for studying card shuffling: these Markov chains are of intrinsic mathematical interest, they model important physical processes in which the positions of particles are interchanged, and they can also serve as simplified models for large-scale mutations—see Section 16.2.

8.1. The Symmetric Group

A stack of n cards can be viewed as an element of the symmetric group $\mathcal{S}_n$ consisting of all permutations of the standard n-element set $\{1, 2, \dots, n\}$. This set forms a group under the operation of functional composition. The identity element of $\mathcal{S}_n$ is the identity function $\mathrm{id}(k) = k$. Every $\sigma \in \mathcal{S}_n$ has a well-defined inverse function, which is its inverse in the group.

A probability distribution μ on the symmetric group describes a mechanism for shuffling cards: apply permutation σ to the deck with probability $\mu(\sigma)$. Repeatedly shuffling the deck using this mechanism is equivalent to running the random walk on the group with increment distribution μ. As discussed in Section 2.6, as long as the support of μ generates all of $\mathcal{S}_n$, the resulting chain is irreducible. If $\mu(\mathrm{id}) > 0$, then it is aperiodic. Every shuffle chain has uniform stationary distribution.

It is most natural to interpret permutations as acting on the locations of cards, rather than their values, and we will do so throughout this chapter. For example, the permutation $\sigma \in S_4$ for which we have

i	1	2	3	4
$\sigma(i)$	3	1	2	4

corresponds to inserting the top card (card 1) into position 3, which pushes card 2 into position 1 and card 3 into position 2 while leaving card 4 fixed.

8.1.1. Cycle notation. We will often find it convenient to use ***cycle notation*** for permutations. In this notation, (abc) refers to the permutation σ for which $b = \sigma(a)$, $c = \sigma(b)$, and $a = \sigma(c)$. When several cycles are written consecutively, they are performed one at a time, *from right to left* (as is consistent with ordinary function composition). For example,

$$(13)(12) = (123) \tag{8.1}$$

and

$$(12)(23)(34)(23)(12) = (14).$$

A cycle of length n is called an n-cycle. A ***transposition*** is a 2-cycle.

In card language, (8.1) corresponds to first exchanging the top and second cards and then interchanging the top and third cards. The result is to send the top card to the second position, the second card to the third position, and the third card to the top of the deck.

Every permutation can be written as a product of disjoint cycles. Fixed points correspond to 1-cycles, which are generally omitted from the notation.

8.1.2. Generating random permutations. We describe a simple algorithm for generating an *exactly* uniform random permutation. Let σ_0 be the identity permutation. For $k = 1, 2, \ldots, n-1$ inductively construct σ_k from σ_{k-1} by swapping the cards at locations k and J_k, where J_k is an integer picked uniformly in $\{k, \ldots, n\}$, independently of $\{J_1, \ldots, J_{k-1}\}$. More precisely,

$$\sigma_k(i) = \begin{cases} \sigma_{k-1}(i) & \text{if } i \neq J_k,\ i \neq k, \\ \sigma_{k-1}(J_k) & \text{if } i = k, \\ \sigma_{k-1}(k) & \text{if } i = J_k. \end{cases} \tag{8.2}$$

Exercise 8.1 asks you to prove that this generates a uniformly chosen element of $\mathcal{S}_n$.

This method requires n steps, which is quite efficient. However, this is not how any human being shuffles cards! In Section 8.3 we will examine a model which comes closer to modeling actual human shuffles.

8.1.3. Parity of permutations. Given a permutation $\sigma \in \mathcal{S}_n$, consider the sign of the product

$$M(\sigma) = \prod_{1 \leq i < j \leq n} (\sigma(j) - \sigma(i)).$$

Clearly $M(\text{id}) > 0$, since every term is positive. For every $\sigma \in \mathcal{S}_n$ and every transposition (ab), we have

$$M((ab)\sigma) = -M(\sigma).$$

Why? We may assume that $a < b$. Then for every c such that $a < c < b$, two factors change sign (the one that pairs c with a and also the one that pairs c with b), while the single factor containing both a and b also changes sign.

Call a permutation σ ***even*** if $M(\sigma) > 0$, and otherwise call σ ***odd***. Note that a permutation is even (odd) if and only if every way of writing it as a product of transpositions contains an even (odd) number of factors. Furthermore, under composition of permutations, evenness and oddness follow the same rules as they do for integer addition. Hence the set of all even permutations in S_n forms a subgroup, known as the ***alternating group*** A_n.

Note that an m-cycle can be written as a product of $m-1$ transpositions:

$$(a_1 a_2 \dots a_n) = (a_1 a_2)(a_2 a_3) \dots (a_{n-1} a_n).$$

Hence an m-cycle is odd (even) when m is even (odd), and the sign of any permutation is determined by its disjoint cycle decomposition.

EXAMPLE 8.1 (Random 3-cycles). Let T be the set of all three-cycles in $\mathcal{S}_n$, and let μ be uniform on T. The set T does *not* generate all of $\mathcal{S}_n$, since every permutation in T is even. Hence the random walk with increments μ is not irreducible. (See Exercise 8.2.)

EXAMPLE 8.2 (Random transpositions, first version). Let $T \subseteq \mathcal{S}_n$ be the set of all transpositions and let μ be the uniform probability distribution on T. In Section 8.1.2, we gave a method for generating a uniform random permutation that started with the identity permutation and used only transpositions. Hence $\langle T \rangle = \mathcal{S}_n$, and our random walk is irreducible.

Every element of the support of μ is odd. Hence, if this walk is started at the identity, after an even number of steps, its position must be an even permutation. After an odd number of steps, its position must be odd. Hence the walk is periodic.

REMARK 8.3. Periodicity occurs in random walks on groups when the entire support of the increment distribution falls into a single coset of some subgroup. Fortunately, there is a simple way to assure aperiodicity. When the probability distribution μ on a group G satisfies $\mu(\mathrm{id}) > 0$, then the random walk with increment distribution μ is aperiodic.

Why? Let $g \in G$. Since $\mu(\mathrm{id}) = P(g, \mathrm{id} \cdot g) = P(g, g) > 0$, we have $1 \in \{t : P^t(g,g) > 0\}$ and thus $\gcd\{t : P^t(g,g) > 0\} = 1$.

EXAMPLE 8.4 (Lazy random transpositions). There is a natural way to modify the random transpositions walk that uses the trick of Remark 8.3 to achieve aperiodicity. At time t, the shuffler chooses two cards, L_t and R_t, independently and uniformly at random. If L_t and R_t are different, transpose them. Otherwise, do nothing. The resulting distribution μ satisfies

$$\mu(\sigma) = \begin{cases} 1/n & \text{if } \rho = \mathrm{id}, \\ 2/n^2 & \text{if } \rho = (ij), \\ 0 & \text{otherwise.} \end{cases} \tag{8.3}$$

8.2. Random Transpositions

It is difficult to imagine a simpler shuffle than the version of random transpositions given in Example 8.4. How many random transpositions are necessary before the deck has been well-randomized?

In Section 8.1.2, we gave a method for generating a uniform random permutation that started with the set $[n]$ sorted and used only transpositions. Thus the set of transpositions generates S_n and by Proposition 2.13 the underlying Markov chain is therefore irreducible.

In each round of random transposition shuffling, (almost) two cards are selected, and each is moved to an almost uniformly random location. In other examples, such as the hypercube, we have been able to bound convergence by tracking how many features have been randomized. If a similar analysis applies to the random transposition shuffle, we might hope that, since each step moves (almost) two cards,

Aligning one card:

$$\begin{matrix} 2 & 4 & 1 & 3 \\ 3 & 1 & 4 & 2 \end{matrix} \Longrightarrow \begin{matrix} 1 & 4 & 2 & 3 \\ 1 & 3 & 4 & 2 \end{matrix}$$

Aligning two cards:

$$\begin{matrix} 2 & 3 & 1 & 4 \\ 3 & 1 & 4 & 2 \end{matrix} \Longrightarrow \begin{matrix} 1 & 3 & 2 & 4 \\ 1 & 3 & 4 & 2 \end{matrix}$$

Aligning three cards:

$$\begin{matrix} 2 & 3 & 1 \\ 3 & 1 & 2 \end{matrix} \Longrightarrow \begin{matrix} 1 & 3 & 2 \\ 1 & 3 & 2 \end{matrix}$$

FIGURE 8.1. Aligning cards using coupled random transpositions. In each example, $X_t = 1$ and $Y_t = 1$, so card 1 is transposed with the card in position 1 in both decks.

half the coupon collector time of approximately $n \log n$ steps will suffice to bring the distribution close to uniform.

In fact, as Diaconis and Shahshahani (1981) proved, the random transpositions walk has a sharp cutoff (see Chapter 18) of width $O(n)$ at $(1/2)n \log n$. They use Fourier analysis on the symmetric group to achieve these extremely precise results. Here, we present two upper bounds on the mixing time: a simple coupling that gives an upper bound of order n^2 for the mixing time and a strong stationary time argument due to Broder (see Diaconis (1988)) that gives an upper bound within a constant factor of the asymptotically sharp answer. While the lower bound we give does not quite reach the cutoff, it does have the correct lead term constant.

8.2.1. Upper bound via coupling. For the coupling, we take a slightly different view of generating the transpositions. At each time t, the shuffler chooses a card $X_t \in [n]$ and, independently, a position $Y_t \in [n]$; she then transposes the card X_t with the card in position Y_t. Of course, if X_t already occupies Y_t, the deck is left unchanged. Hence this mechanism generates the distribution described in (8.3).

To couple two decks, use the same choices (X_t) and (Y_t) to shuffle both. Let (σ_t) and (σ'_t) be the two trajectories. What can happen in one step? Let a_t be the number of cards that occupy the same position in both σ_t and σ'_t.

- If X_t is in the same position in both decks and the same card occupies position Y_t in both decks, then $a_{t+1} = a_t$.
- If X_t is in different positions in the two decks but position Y_t is occupied by the same card, then performing the specified transposition breaks one alignment but also forms a new one. We have $a_{t+1} = a_t$.
- If X_t is in different positions in the two decks and if the cards at position Y_t in the two decks do not match, then at least one new alignment is made—and possibly as many as three. See Figure 8.1.

PROPOSITION 8.5. *Let τ be the time required for the two decks to couple. Then, no matter the initial configurations of the two decks, $\mathbf{E}(\tau) < \frac{\pi^2}{6}n^2$.*

PROOF. Decompose

$$\tau = \tau_1 + \cdots + \tau_n,$$

where τ_i is the number of transpositions between the first time that a_t is greater than or equal to $i-1$ and the first time that a_t is greater than or equal to i. (Since a_0 can be greater than 0 and since a_t can increase by more than 1 in a single transposition, it is possible that many of the τ_i's are equal to 0.)

When t satisfies $a_t = i$, there are $n-i$ unaligned cards and the probability of increasing the number of alignments is $(n-i)^2/n^2$, since the shuffler must choose a non-aligned card and a non-aligned position. In this situation τ_{i+1} is a geometric random variable with success probability $(n-i)^2/n^2$. We may conclude that under these circumstances

$$\mathbf{E}(\tau_{i+1}|a_t = i) = n^2/(n-i)^2.$$

When no value of t satisfies $a_t = i$, then $\tau_{i+1} = 0$. Hence

$$\mathbf{E}(\tau) < n^2 \sum_{i=0}^{n-1} \frac{1}{(n-i)^2} < n^2 \sum_{l=1}^{\infty} \frac{1}{l^2}.$$

■

Markov's inequality and Corollary 5.3 now give an $O(n^2)$ bound on t_{mix}. However, the strong stationary time we are about to discuss does much better.

8.2.2. Upper bound via strong stationary time.

PROPOSITION 8.6. *In the random transposition shuffle, let R_t and L_t be the cards chosen by the right and left hands, respectively, at time t. Assume that when $t = 0$, no cards have been marked. At time t, mark card R_t if both of the following are true:*

- *R_t is unmarked.*
- *Either L_t is a marked card or $L_t = R_t$.*

Let τ be the time when every card has been marked. Then τ is a strong stationary time for this chain.

Here is a heuristic explanation for why the scheme described above should give a strong stationary time. One way to generate a uniform random permutation is to build a stack of cards, one at a time, inserting each card into a uniformly random position relative to the cards already in the stack. For the stopping time described above, the marked cards are carrying out such a process.

PROOF. It is clear that τ is a stopping time. To show that it is a strong stationary time, we prove the following subclaim by induction on t. Let $V_t \subseteq [n]$ be the set of cards marked at or before time t, and let $U_t \subseteq [n]$ be the set of positions occupied by V_t after the t-th transposition. We claim that *given t, V_t, and U_t, all possible permutations of the cards in V_t on the positions U_t are equally likely.*

This is clearly true when $t = 1$ (and continues to clearly be true as long as at most one card has been marked).

Now, assume that the subclaim is true for t. The shuffler chooses cards L_{t+1} and R_{t+1}.

- If no new card is marked, then $V_{t+1} = V_t$. This can happen in three ways:
  - The cards L_{t+1} and R_{t+1} are different and both are unmarked. Then V_{t+1} and U_{t+1} are identical to V_t and U_t, respectively.

- If L_{t+1} and R_{t+1} were both marked at an earlier round, then $U_{t+1} = U_t$ and the shuffler applies a uniform random transposition to the cards in V_t. All permutations of V_t remain equiprobable.
- Otherwise, L_{t+1} is unmarked and R_{t+1} was marked at an earlier round. To obtain the position set U_{t+1}, we delete the position (at time t) of R_{t+1} and add the position (at time t) of L_{t+1}. For a fixed set U_t, all choices of $R_{t+1} \in U_t$ are equally likely, as are all permutations of V_t on U_t. Hence, once the positions added and deleted are specified, all permutations of V_t on U_{t+1} are equally likely.

- If the card R_{t+1} gets marked, then L_{t+1} is equally likely to be any element of $V_{t+1} = V_t \cup \{R_{t+1}\}$, while U_{t+1} consists of U_t along with the position of R_{t+1} (at time t). Specifying the permutation of V_t on U_t and the card L_{t+1} uniquely determines the permutation of V_{t+1} on U_{t+1}. Hence all such permutations are equally likely.

In every case, the collection of all permutations of the cards V_t on a specified set U_t together make equal contributions to all possible permutations of V_{t+1} on U_{t+1}. Hence, to conclude that all possible permutations of a fixed V_{t+1} on a fixed U_{t+1} are equally likely, we simply sum over all possible preceding configurations. ■

REMARK 8.7. In the preceding proof, the two subcases of the inductive step for which no new card is marked are essentially the same as checking that the uniform distribution is stationary for the random transposition shuffle and the random-to-top shuffle, respectively.

REMARK 8.8. As Diaconis (1988) points out, for random transpositions some simple card-marking rules fail to give strong stationary times. See Exercise 8.8.

LEMMA 8.9. *The stopping time τ defined in Proposition 8.6 satisfies*

$$\mathbf{E}(\tau) = 2n(\log n + O(1))$$

and

$$\mathrm{Var}(\tau) = O(n^2).$$

PROOF. As for the coupon collector time, we can decompose

$$\tau = \tau_0 + \cdots + \tau_{n-1},$$

where τ_k is the number of transpositions after the k-th card is marked, up to and including when the $(k+1)$-st card is marked. The rules specified in Proposition 8.6 imply that τ_k is a geometric random variable with success probability $\frac{(k+1)(n-k)}{n^2}$ and that the τ_i's are independent of each other. Hence

$$\mathbf{E}(\tau) = \sum_{k=0}^{n-1} \frac{n^2}{(k+1)(n-k)}.$$

Substituting the partial fraction decomposition

$$\frac{1}{(k+1)(n-k)} = \frac{1}{n+1}\left(\frac{1}{k+1} + \frac{1}{n-k}\right)$$

and recalling that

$$\sum_{j=1}^{n} \frac{1}{j} = \log n + O(1)$$

(see Exercise 2.4) completes the estimate.

Now, for the variance. We can immediately write

$$\mathrm{Var}(\tau) = \sum_{k=0}^{n-1} \frac{1 - \frac{(k+1)(n-k)}{n^2}}{\left(\frac{(k+1)(n-k)}{n^2}\right)^2} < \sum_{k=0}^{n-1} \frac{n^4}{(k+1)^2(n-k)^2}.$$

Split the sum into two pieces:

$$\begin{aligned}\mathrm{Var}(\tau) &< \sum_{0 \le k < n/2} \frac{n^4}{(k+1)^2(n-k)^2} + \sum_{n/2 \le k < n} \frac{n^4}{(k+1)^2(n-k)^2} \\ &< \frac{2n^4}{(n/2)^2} \sum_{0 \le k \le n/2} \frac{1}{(k+1)^2} = O(n^2).\end{aligned}$$

■

COROLLARY 8.10. *For the random transposition chain on an n-card deck,*

$$t_{\mathrm{mix}} \le (2 + o(1)) n \log n.$$

PROOF. Let τ be the Broder stopping time defined in Proposition 8.6, and let $t_0 = E(\tau) + 2\sqrt{\mathrm{Var}(\tau)}$. By Chebyshev's inequality,

$$\mathbf{P}\{\tau > t_0\} \le \frac{1}{4}.$$

Lemma 8.9 and Proposition 6.10 now imply the desired inequality. ■

8.2.3. Lower bound.

PROPOSITION 8.11. *Let $0 < \varepsilon < 1$. For the random transposition chain on an n-card deck,*

$$t_{\mathrm{mix}}(\varepsilon) \ge \frac{n-1}{2} \log\left(\frac{1-\varepsilon}{6} n\right).$$

PROOF. It is well known (and easily proved using indicators) that the expected number of fixed points in a uniform random permutation in $\mathcal{S}_n$ is 1, regardless of the value of n.

Let $F(\sigma)$ denote the number of fixed points of the permutation σ. If σ is obtained from the identity by applying t random transpositions, then $F(\sigma)$ is at least as large as the number of cards that were touched by none of the transpositions—no such card has moved, and some moved cards may have returned to their original positions.

Our shuffle chain determines transpositions by choosing pairs of cards independently and uniformly at random. Hence, after t shuffles, the number of untouched cards has the same distribution as the number R_{2t} of uncollected coupon types after $2t$ steps of the coupon collector chain.

Let $A = \{\sigma : F(\sigma) \ge \mu/2\}$. We will compare the probabilities of A under the uniform distribution π and $P^t(\mathrm{id}, \cdot)$. First,

$$\pi(A) \le \frac{2}{\mu},$$

by Markov's inequality. On the other hand, by Lemma 7.12, R_{2t} has expectation

$$\mu = n\left(1 - \frac{1}{n}\right)^{2t}$$

and variance bounded by μ. By Chebyshev,

$$P^t(\mathrm{id}, A^c) \le \mathbf{P}\{R_{2t} \le \mu/2\} \le \frac{\mu}{(\mu/2)^2} = \frac{4}{\mu}.$$

By the definition (4.1) of total variation distance, we have

$$\left\|P_n^t(\mathrm{id}, \cdot) - \pi\right\|_{\mathrm{TV}} \ge 1 - \frac{6}{\mu}.$$

We want to find how small t must be so that $1 - 6/\mu \ge \varepsilon$, or equivalently,

$$n\left(1 - \frac{1}{n}\right)^{2t} = \mu \ge \frac{6}{1-\varepsilon}.$$

The above holds if and only if

$$\log\left(\frac{n(1-\varepsilon)}{6}\right) \ge 2t \log\left(\frac{n}{n-1}\right). \tag{8.4}$$

Using the inequality $\log(1+x) \le x$, we have $\log\left(\frac{n}{n-1}\right) \le \frac{1}{n-1}$, so the inequality (8.4) holds provided that

$$\log\left(\frac{n(1-\varepsilon)}{6}\right) \ge \frac{2t}{n-1}.$$

That is, if $t \le \frac{n-1}{2}\log\left(\frac{n(1-\varepsilon)}{6}\right)$, then $d(t) \ge 1 - 6/\mu \ge \varepsilon$. ∎

8.3. Riffle Shuffles

The method most often used to shuffle real decks of 52 cards is the following: first, the shuffler cuts the decks into two piles. Then, the piles are "riffled" together: she successively drops cards from the bottom of each pile to form a new pile. There are two undetermined aspects of this procedure. First, the numbers of cards in each pile after the initial cut can vary. Second, real shufflers drop varying numbers of cards from each stack as the deck is reassembled.

Fortunately for mathematicians, there is a tractable mathematical model for riffle shuffling. Here are three ways to shuffle a deck of n cards:

(1) Let M be a binomial$(n, 1/2)$ random variable, and split the deck into its top M cards and its bottom $n - M$ cards. There are $\binom{n}{M}$ ways to riffle these two piles together, preserving the relative order within each pile (first select the positions for the top M cards; then fill in both piles). Choose one of these arrangements uniformly at random.

(2) Let M be a binomial$(n, 1/2)$ random variable, and split the deck into its top M cards and its bottom $n - M$ cards. The two piles are then held over the table and cards are dropped one by one, forming a single pile once more, according to the following recipe: if at a particular moment, the left pile contains a cards and the right pile contains b cards, then drop the card on the bottom of the left pile with probability $a/(a+b)$ and the card on the bottom of the right pile with probability $b/(a+b)$. Repeat this procedure until all cards have been dropped.

(3) Label the n cards with n independent fairly chosen bits. Pull all the cards labeled 0 to the top of the deck, preserving their relative order.

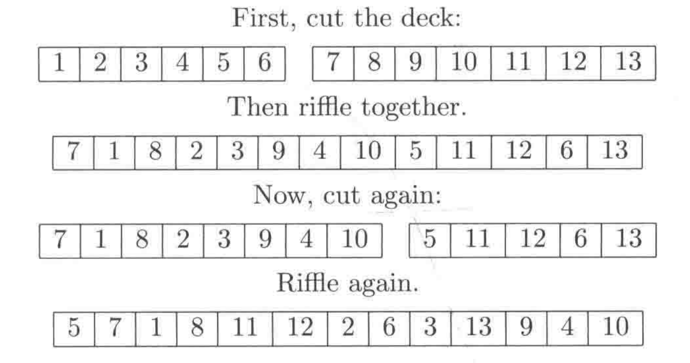


FIGURE 8.2. Riffle shuffling a 13-card deck, twice.

A ***rising sequence*** of a permutation σ is a maximal set of consecutive values that occur in the correct relative order in σ. (For example, the final permutation in Figure 8.2 has 4 rising sequences: $(1, 2, 3, 4), (5, 6), (7, 8, 9, 10)$, and $(11, 12, 13)$.

We claim that *methods* (1) *and* (2) *generate the same distribution* Q *on permutations, where*

$$Q(\sigma) = \begin{cases} (n+1)/2^n & \text{if } \sigma = \text{id}, \\ 1/2^n & \text{if } \sigma \text{ has exactly two rising sequences}, \\ 0 & \text{otherwise.} \end{cases} \tag{8.5}$$

It should be clear that method (1) generates Q; the only tricky detail is that the identity permutation is always an option, no matter the value of M. Given M, method (2) assigns probability $M!(n-M)!/n! = \binom{n}{M}^{-1}$ to each possible interleaving, since each step drops a single card and every card must be dropped.

Recall from Section 4.6 that for a distribution R on $\mathcal{S}_n$, the ***inverse distribution*** $\widehat{R}$ satisfies $\widehat{R}(\rho) = R(\rho^{-1})$. We claim that *method* (3) *generates* $\widehat{Q}$. Why? The cards labeled 0 form one increasing sequence in ρ^{-1}, and the cards labeled 1 form the other. (Again, there are $n+1$ ways to get the identity permutation, namely, all strings of the form $00\dots011\dots1$.)

Thanks to Lemma 4.13 (which says that a random walk on a group and its inverse, both started from the identity, have the same distance from uniformity after the same number of steps), it will suffice to analyze method (3).

Now, consider repeatedly inverse riffle shuffling a deck, using method (3). For the first shuffle, each card is assigned a random bit, and all the 0's are pulled ahead of all the 1's. For the second shuffle, each card is again assigned a random bit, and all the 0's are pulled ahead of all the 1's. Considering both bits (and writing the second bit on the left), we see that cards labeled 00 precede those labeled 01, which precede those labeled 10, which precede those labeled 11 (see Figure 8.3). After k shuffles, each card will be labeled with a string of k bits, and cards with different labels will be in lexicographic order (cards with the same label will be in their original relative order).

PROPOSITION 8.12. *Let τ be the number of inverse riffle shuffles required for all cards to have different bitstring labels. Then τ is a strong stationary time.*

PROOF. Assume $\tau = t$. Since the bitstrings are generated by independent fair coin flips, every assignment of strings of length t to cards is equally likely. Since the

Initial order:

card	1	2	3	4	5	6	7	8	9	10	11	12	13
round 1	1	0	0	1	1	1	0	1	0	1	1	0	0
round 2	0	1	0	1	0	1	1	1	0	0	1	0	1

After one inverse riffle shuffle:

card	2	3	7	9	12	13	1	4	5	6	8	10	11
round 1	0	0	0	0	0	0	1	1	1	1	1	1	1
round 2	1	0	1	0	0	1	0	1	0	1	1	0	1

After two inverse riffle shuffles:

card	3	9	12	1	5	10	2	7	13	4	6	8	11
round 1	0	0	0	1	1	1	0	0	0	1	1	1	1
round 2	0	0	0	0	0	0	1	1	1	1	1	1	1

FIGURE 8.3. When inverse riffle shuffling, we first assign bits for each round, then sort bit by bit.

labeling bitstrings are distinct, the permutation is fully determined by the labels. Hence the permutation of the cards at time τ is uniform, no matter the value of τ. ■

Now we need only estimate the tail probabilities for the strong stationary time. However, our stopping time τ is an example of the birthday problem, with the slight twist that the number of "people" is fixed, and we wish to choose an appropriate power-of-two "year length" so that all the people will, with high probability, have different birthdays.

PROPOSITION 8.13. *For the riffle shuffle on an n-card deck, $t_{\mathrm{mix}} \leq 2\log_2(4n/3)$ for sufficiently large n.*

PROOF. Consider inverse riffle shuffling an n-card deck and let τ be the stopping time defined in Proposition 8.12. If $\tau \leq t$, then different labels have been assigned to all n cards after t inverse riffle shuffles. Hence

$$\mathbf{P}(\tau \leq t) = \prod_{k=0}^{n-1}\left(1 - \frac{k}{2^t}\right),$$

since there are 2^t possible labels. Let $t = 2\log_2(n/c)$. Then $2^t = n^2/c^2$ and we have

$$\begin{aligned}\log \prod_{k=0}^{n-1}\left(1 - \frac{k}{2^t}\right) &= -\sum_{k=0}^{n-1}\left(\frac{c^2k}{n^2} + O\left(\frac{k}{n^2}\right)^2\right) \\ &= -\frac{c^2n(n-1)}{2n^2} + O\left(\frac{n^3}{n^4}\right) = -\frac{c^2}{2} + O\left(\frac{1}{n}\right).\end{aligned}$$

Hence

$$\lim_{n\to\infty} \frac{\mathbf{P}(\tau \leq t)}{e^{-c^2/2}} = 1.$$

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

FIGURE 8.4. The "fifteen puzzle".

Taking any value of c such that $c < \sqrt{2\log(4/3)} \approx 0.7585$ will give a bound on $t_{\rm mix} = t_{\rm mix}(1/4)$. A convenient value to use is $3/4$, which, combined with Proposition 6.10, gives the bound stated in the proposition. ■

Applying the counting bound in Section 7.1.1 gives a lower bound of logarithmic order on the mixing time for the riffle shuffle.

PROPOSITION 8.14. *Fix $0 < \varepsilon, \delta < 1$. Consider riffle shuffling an n-card deck. For sufficiently large n,*

$$t_{\rm mix}(\varepsilon) \geq (1-\delta)\log_2 n. \tag{8.6}$$

PROOF. There are at most 2^n possible states accessible in one step of the chain, since we can generate a move using n independent unbiased bits. Thus $\log_2 \Delta \leq n$, where Δ is the maximum out-degree defined in (7.1). The state space has size $n!$, and Stirling's formula shows that $\log_2 n! = [1+o(1)]n\log_2 n$. Using these estimates in (7.2) shows that for all $\delta > 0$, if n is sufficiently large then (8.6) holds. ■

Exercises

EXERCISE 8.1. Let $J_1, \ldots, J_{n-1}$ be independent integers, where J_k is uniform on $\{k, k+1, \ldots, n\}$, and let σ_{n-1} be the random permutation obtained by recursively applying (8.2). Show that σ_{n-1} is uniformly distributed on $\mathcal{S}_n$.

EXERCISE 8.2.

(a) Show that the alternating group $A_n \subseteq \mathcal{S}_n$ of even permutations has order $n!/2$.
(b) Consider the distribution μ, uniform on the set of 3-cycles in $\mathcal{S}_n$, introduced in Example 8.1. Show that the random walk with increments μ is an irreducible and aperiodic chain when considered as a walk on A_n.

EXERCISE 8.3. The long-notorious Sam Loyd "fifteen puzzle" is shown in Figure 8.4. It consists of 15 tiles, numbered with the values 1 through 15, sitting in a 4 by 4 grid; one space is left empty. The tiles are in order, except that tiles 14 and 15 have been switched. The only allowed moves are to slide a tile adjacent to the empty space into the empty space.

Is it possible, using only legal moves, to switch the positions of tiles 14 and 15, while leaving the rest of the tiles fixed?

(a) Show that the answer is "no."
(b) Describe the set of all configurations of tiles that can be reached using only legal moves.

EXERCISE 8.4. Suppose that a random function $\sigma : [n] \to [n]$ is created by letting $\sigma(i)$ be a random element of $[n]$, independently for each $i = 1, \ldots, n$. If the resulting function σ is a permutation, stop, and otherwise begin anew by generating a fresh random function. Use Stirling's formula to estimate the expected number of random functions generated up to and including the first permutation.

EXERCISE 8.5. Consider the following variation of our method for generating random permutations: let σ_0 be the identity permutation. For $k = 1, 2, \ldots, n$ inductively construct σ_k from σ_{k-1} by swapping the cards at locations k and J_k, where J_k is an integer picked uniformly in $[1, n]$, independently of previous picks.

For which values of n does this variant procedure yield a uniform random permutation?

EXERCISE 8.6. True or false: let Q be a distribution on $\mathcal{S}_n$ such that when $\sigma \in \mathcal{S}_n$ is chosen according to Q, we have

$$\mathbf{P}(\sigma(i) > \sigma(j)) = 1/2$$

for every $i, j \in [n]$. Then Q is uniform on $\mathcal{S}_n$.

EXERCISE 8.7. Kolata (January 9, 1990) writes: "By saying that the deck is completely mixed after seven shuffles, Dr. Diaconis and Dr. Bayer mean that every arrangement of the 52 cards is equally likely or that any card is as likely to be in one place as in another."

True or false: let Q be a distribution on $\mathcal{S}_n$ such that when $\sigma \in \mathcal{S}_n$ is chosen according to Q, we have

$$\mathbf{P}(\sigma(i) = j) = 1/n$$

for every $i, j \in [n]$. Then Q is uniform on $\mathcal{S}_n$.

EXERCISE 8.8. Consider the random transposition shuffle.

(a) Show that marking both cards of every transposition and proceeding until every card is marked does not yield a strong stationary time.
(b) Show that marking the right-hand card of every transposition and proceeding until every card is marked does not yield a strong stationary time.

EXERCISE 8.9. Let $\varphi : [n] \to \mathbb{R}$ be any function. Let $\sigma \in \mathcal{S}_n$. Show that the value of

$$\varphi_\sigma = \sum_{k \in [n]} \varphi(k)\varphi(\sigma(k))$$

is maximized when $\sigma = \mathrm{id}$.

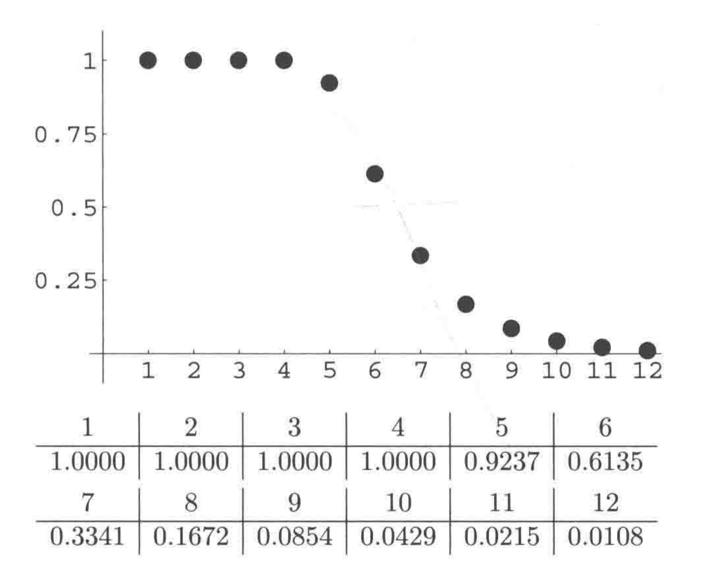
EXERCISE 8.10. Show that for any positive integer n,

$$\sum_{k \in [n]} \cos^2\left(\frac{(2k-1)\pi}{2n}\right) = \frac{n}{2}.$$

EXERCISE 8.11. Here is a way to generalize the inverse riffle shuffle. Let a be a positive integer. To perform an ***inverse a-shuffle***, assign independent uniform random digits chosen from $\{0, 1, \ldots, a-1\}$ to each card. Then sort according to digit, preserving relative order for cards with the same digit. For example, if $a = 3$ and the digits assigned to cards are

1	2	3	4	5	6	7	8	9	10	11	12
2	0	2	1	2	0	1	0	1	0	0	0

,



1	2	3	4	5	6
1.0000	1.0000	1.0000	1.0000	0.9237	0.6135
7	**8**	**9**	**10**	**11**	**12**
0.3341	0.1672	0.0854	0.0429	0.0215	0.0108

FIGURE 8.5. The total variation distance from stationarity (with 4 digits of precision) after t riffle shuffles of a 52-card deck, for $t = 1, \ldots, 12$.

then the shuffle will give

$$2 \mid 6 \mid 8 \mid 10 \mid 11 \mid 12 \mid 4 \mid 7 \mid 9 \mid 1 \mid 3 \mid 5 \ .$$

(a) Let a and b be positive integers. Show that an inverse a-shuffle followed by an inverse b-shuffle is the same as an inverse ab-shuffle.

(b) Describe (mathematically) how to perform a *forwards* a-shuffle, and show that its increment distribution gives weight $\binom{a+n-r}{n}/a^n$ to every $\sigma \in \mathcal{S}_n$ with exactly r rising sequences. (This is a generalization of (8.5).)

REMARK 8.15. Exercise 8.11(b), due to Bayer and Diaconis (1992), is the key to numerically computing the total variation distance from stationarity. A permutation has r rising sequences if and only if its inverse has $r-1$ descents. The number of permutations in $\mathcal{S}_n$ with $r-1$ descents is the ***Eulerian number*** $\left\langle {n \atop r-1} \right\rangle$. The Eulerian numbers satisfy a simple recursion (and are built into modern symbolic computation software); see p. 267 of Graham, Knuth, and Patashnik (1994) for details. It follows from Exercise 8.11 that the total variation distance from uniformity after t Gilbert-Shannon-Reeds shuffles of an n-card deck is

$$\sum_{r=1}^{n} \left\langle {n \atop r-1} \right\rangle \left| \frac{\binom{2^t+n-r}{n}}{2^{nt}} - \frac{1}{n!} \right| .$$

See Figure 8.5 for the values when $n = 52$ and $t \leq 12$.

Notes

See any undergraduate abstract algebra book, e.g. Herstein (1975) or Artin (1991), for more on the basic structure of the symmetric group $\mathcal{S}_n$.

Thorp (1965) proposed Exercise 8.5 as an "Elementary Problem" in the *American Mathematical Monthly*.

Random transpositions. Our upper bound on the mixing time for random transpositions is off by a factor of 4. Matthews (1988b) gives an improved strong stationary time whose upper bound matches the lower bound. Here is how it works: again, let R_t and L_t be the cards chosen by the right and left hands, respectively, at time t. Assume that when $t = 0$, no cards have been marked. As long as at most $\lceil n/3 \rceil$ cards have been marked, use this rule: at time t, mark card R_t if both R_t and L_t are unmarked. When $k > \lceil n/3 \rceil$ cards have been marked, the rule is more complicated. Let $l_1 < l_2 < \cdots < l_k$ be the marked cards, and enumerate the ordered pairs of marked cards in lexicographic order:

$$(l_1, l_1), (l_1, l_2), \ldots, (l_1, l_k), (l_2, l_1), \ldots, (l_k, l_k). \tag{8.7}$$

Also list the unmarked cards in order: $u_1 < u_n < \cdots < u_{n-k}$. At time t, if there exists an i such that $1 \leq i \leq n-k$ and one of the three conditions below is satisfied, then mark card i.

(i) $L_t = R_t = u_i$.
(ii) Either $L_t = u_i$ and R_t is marked or $R_t = u_i$ and L_t is marked.
(iii) The pair (L_t, R_t) is identical to the i-th pair in the list (8.7) of pairs of marked cards.

(Note that at most one card can be marked per transposition; if case (iii) is invoked, the card marked may not be either of the selected cards.) Compared to the Broder time discussed earlier, this procedure marks cards much faster at the beginning and essentially twice as fast at the end. The analysis is similar in spirit to, but more complex than, that presented in Section 8.2.2.

Semi-random transpositions. Consider shuffling by transposing cards. However, we allow only one hand (the right) to choose a uniform random card. The left hand picks a card according to some other rule—perhaps deterministic, perhaps randomized—and the two cards are switched. Since only one of the two cards switched is fully random, it is reasonable to call examples of this type shuffles by ***semi-random transpositions***. (Note that for this type of shuffle, the distribution of allowed moves can depend on time.)

One particularly interesting variation first proposed by Thorp (1965) and mentioned as an open problem in Aldous and Diaconis (1986) is the ***cyclic-to-random*** shuffle: at step t, the left hand chooses card $t \pmod n$, the right hand chooses a uniform random card, and the two chosen cards are transposed. This chain has the property that every position is given a chance to be randomized once every n steps. Might that speed randomization? Or does the reduced randomness slow it down? (Note: Exercise 8.5 is about the state of an n-card deck after n rounds of cyclic-to-random transpositions.)

Mironov (2002) (who was interested in how many steps are needed to do a good job of initializing a standard cryptographic protocol) gives an $O(n \log n)$ upper bound, using a variation of Broder's stopping time for random transpositions. Mossel, Peres, and Sinclair (2004) prove a matching (to within a constant) lower bound. Furthermore, the same authors extend the stopping time argument to give an $O(n \log n)$ upper bound for *any* shuffle by semi-random transpositions. See also Ganapathy (2007).

Riffle shuffles. The most famous theorem in non-asymptotic Markov chain convergence is what is often, and perhaps unfortunately, called the "seven shuffles

suffice" (for mixing a standard 52-card deck) result of Bayer and Diaconis (1992), which was featured in the New York Times (Kolata, January 9, 1990). Many elementary expositions of the riffle shuffle have been written. Our account is in debt to Aldous and Diaconis (1986), Diaconis (1988), and Mann (1994).

The model for riffle shuffling that we have discussed was developed by Gilbert and Shannon at Bell Labs in the 1950's and later independently by Reeds. It is natural to ask whether the Gilbert-Shannon-Reeds (GSR) shuffle is a reasonable model for the way humans riffle cards together. Diaconis (1988) reports that when he and Reeds both shuffled repeatedly, Reeds's shuffles had packet sizes that matched the GSR model well, while Diaconis's shuffles had more small packets. The difference is not surprising, since Diaconis is an expert card magician who can perform perfect shuffles—i.e., ones in which a single card is dropped at a time.

Far more is known about the GSR shuffle than we have discussed. Bayer and Diaconis (1992) derived the exact expression for the probability of any particular permutation after t riffle shuffles discussed in Exercise 8.11 and showed that the riffle shuffle has a cutoff (in the sense we discuss in Chapter 18) when $t = \frac{3}{2} n \log n$. Diaconis, McGrath, and Pitman (1995) compute exact probabilities of various properties of the resulting permutations and draw beautiful connections with combinatorics and algebra. See Diaconis (2003) for a survey of mathematics that has grown out of the analysis of the riffle shuffle.

Is it in fact true that seven shuffles suffice to adequately randomize a 52-card deck? Bayer and Diaconis (1992) were the first to give explicit values for the total variation distance from stationarity after various numbers of shuffles; see Figure 8.5. After seven shuffles, the total variation distance from stationarity is approximately 0.3341. That is, after 7 riffle shuffles the probability of a given event can differ by as much as 0.3341 from its value under the uniform distribution. Indeed, Peter Doyle has described a simple solitaire game for which the probability of winning when playing with a uniform random deck is exactly 1/2, but whose probability of winning with a deck that has been GSR shuffled 7 times from its standard order is 0.801 (as computed in van Zuylen and Schalekamp (2004)).

Ultimately the question of how many shuffles suffice for a 52-card deck is one of opinion, not mathematical fact. However, there exists at least one game playable by human beings for which 7 shuffles clearly do not suffice. A more reasonable level of total variation distance might be around 1 percent, comparable to the house advantage in casino games. This threshold would suggest 11 or 12 as an appropriate number of shuffles.

CHAPTER 9

Random Walks on Networks

9.1. Networks and Reversible Markov Chains

Electrical networks provide a different language for reversible Markov chains. This point of view is useful because of the insight gained from the familiar physical laws of electrical networks.

A ***network*** is a finite undirected connected graph G with vertex set V and edge set E, endowed additionally with non-negative numbers $\{c(e)\}$, called ***conductances***, that are associated to the edges of G. We often write $c(x, y)$ for $c(\{x, y\})$; clearly $c(x, y) = c(y, x)$. The reciprocal $r(e) = 1/c(e)$ is called the ***resistance*** of the edge e.

A network will be denoted by the pair $(G, \{c(e)\})$. Vertices of G are often called ***nodes***. For $x, y \in V$, we will write $x \sim y$ to indicate that $\{x, y\}$ belongs to E.

Consider the Markov chain on the nodes of G with transition matrix

$$P(x, y) = \frac{c(x, y)}{c(x)}, \tag{9.1}$$

where $c(x) = \sum_{y\,:\,y\sim x} c(x, y)$. This process is called the ***weighted random walk*** on G with edge weights $\{c(e)\}$, or the Markov chain associated to the network $(G, \{c(e)\})$. This Markov chain is reversible with respect to the probability π defined by $\pi(x) := c(x)/c_G$, where $c_G = \sum_{x\in V} c(x)$:

$$\pi(x)P(x, y) = \frac{c(x)}{c_G}\frac{c(x, y)}{c(x)} = \frac{c(x, y)}{c_G} = \frac{c(y, x)}{c_G} = \frac{c(y)}{c_G}\frac{c(y, x)}{c(y)} = \pi(y)P(y, x).$$

By Proposition 1.19, π is stationary for P. Note that

$$c_G = \sum_{x\in V}\sum_{\substack{y\in V\\ y\sim x}} c(x, y).$$

In the case that the graph has no loops, we have

$$c_G = 2\sum_{e\in E} c(e).$$

Simple random walk on G, defined in Section 1.4 as the Markov chain with transition probabilities

$$P(x, y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y \sim x, \\ 0 & \text{otherwise}, \end{cases} \tag{9.2}$$

is a special case of a weighted random walk: set the weights of all edges in G equal to 1.

We now show that, in fact, every reversible Markov chain is a weighted random walk on a network. Suppose P is a transition matrix on a finite set Ω which is reversible with respect to the probability π (that is, (1.30) holds). Define a graph

with vertex set Ω by declaring $\{x, y\}$ an edge if $P(x, y) > 0$. This is a proper definition, since reversibility implies that $P(x, y) > 0$ exactly when $P(y, x) > 0$. Next, define conductances on edges by $c(x, y) = \pi(x)P(x, y)$. This is symmetric by reversibility. With this choice of weights, we have $c(x) = \pi(x)$, and thus the transition matrix associated with this network is just P. The study of reversible Markov chains is thus equivalent to the study of random walks on networks.

9.2. Harmonic Functions

We assume throughout this section that P is the transition matrix of an irreducible Markov chain with state space Ω. We do *not* assume in this section that P is reversible; indeed, Proposition 9.1 is true for all irreducible chains.

Recall from Section 1.5.4 that we call a function $h : \Omega \to \mathbb{R}$ ***harmonic*** for P at a vertex x if

$$h(x) = \sum_{y \in \Omega} P(x, y)h(y). \tag{9.3}$$

When P is the transition matrix for a random walk on a graph, (9.3) means that $h(x)$ is the average of the values at h at neighboring vertices.

Recall that when B is a set of states, we define the hitting time τ_B by $\tau_B = \min\{t \geq 0 : X_t \in B\}$.

PROPOSITION 9.1. *Let (X_t) be a Markov chain with irreducible transition matrix P, let $B \subset \Omega$, and let $h_B : B \to \mathbb{R}$ be a function defined on B. The function $h : \Omega \to \mathbb{R}$ defined by $h(x) := \mathbf{E}_x h_B(X_{\tau_B})$ is the unique extension $h : \Omega \to \mathbb{R}$ of h_B such that $h(x) = h_B(x)$ for all $x \in B$ and h is harmonic for P at all $x \in \Omega \setminus B$.*

REMARK 9.2. The proof of uniqueness below, derived from the maximum principle, should remind you of that of Lemma 1.16.

PROOF. We first show that $h(x) = \mathbf{E}_x h(X_{\tau_B})$ is a harmonic extension of h_B. Clearly $h(x) = h_B(x)$ for all $x \in B$. Suppose that $x \in \Omega \setminus B$. Then

$$h(x) = \mathbf{E}_x h(X_{\tau_B}) = \sum_{y \in \Omega} P(x, y)\mathbf{E}_x[h(X_{\tau_B}) \mid X_1 = y]. \tag{9.4}$$

Observe that $x \in \Omega \setminus B$ implies that $\tau_B \geq 1$. By the Markov property, it follows that

$$\mathbf{E}_x[h(X_{\tau_B}) \mid X_1 = y] = \mathbf{E}_y h(X_{\tau_B}) = h(y). \tag{9.5}$$

Substituting (9.5) in (9.4) shows that h is harmonic at x.

We now show uniqueness. Suppose $g : \Omega \to \mathbb{R}$ is a function which is harmonic on $\Omega \setminus B$ and satisfies $g(x) = 0$ for all $x \in B$. We first show that $g \leq 0$. Suppose this is not the case. Let $x \notin B$ belong to the set

$$A := \left\{ x : g(x) = \max_{\Omega \setminus B} g \right\}$$

and suppose that $P(x, y) > 0$. If $g(y) < g(x)$, then harmonicity of g on $\Omega \setminus B$ implies

$$g(x) = \sum_{z \in \Omega} g(z)P(x, z) = g(y)P(x, y) + \sum_{\substack{z \in \Omega \\ z \neq y}} g(z)P(x, z) < \max_{\Omega \setminus B} g,$$

a contradiction. It follows that $g(y) = \max_{\Omega \setminus B} g$, that is, $y \in A$.

By irreducibility, for any $y \in B$, there exists a sequence of states $y_0, y_1, \ldots, y_r$ such that $y_0 = x$ and $y_r = y$ and such that $P(y_{i-1}, y_i) > 0$ for $i = 1, 2, \ldots, r$. Therefore, each $y_i \in A$. In particular, $y \in A$. Since $g(y) = 0$, it follows that $\max_{\Omega \setminus B} g \leq 0$. Since $g(x) = 0$ for $x \in B$, it follows that $\max_{\Omega} g \leq 0$. Applying this argument to $-h$ shows that $\min_{\Omega} g \geq 0$, whence $g(x) = 0$ for all $x \in \Omega$.

Now, if h and $\tilde{h}$ are both harmonic on $\Omega \setminus B$ and agree on B, then the difference $h - \tilde{h}$ is harmonic on $\Omega \setminus B$ and vanishes on B. Therefore, $h(x) - \tilde{h}(x) = 0$ for all $x \in \Omega$. ■

REMARK 9.3. Note that requiring h to be harmonic on $X \setminus B$ yields a system of $|\Omega| - |B|$ linear equations in the $|\Omega| - |B|$ unknowns $\{h(x)\}_{x \in \Omega \setminus B}$. For such a system of equations, existence of a solution implies uniqueness.

9.3. Voltages and Current Flows

Consider a network $(G, \{c(e)\})$. We distinguish two nodes, a and z, which are called the ***source*** and the ***sink*** of the network. A function W which is harmonic on $V \setminus \{a, z\}$ will be called a ***voltage***. Proposition 9.1 implies that a voltage is completely determined by its boundary values $W(a)$ and $W(z)$.

An ***oriented edge*** $\vec{e} = \vec{xy}$ is an *ordered* pair of nodes (x, y). A ***flow*** θ is a function on oriented edges which is antisymmetric, meaning that $\theta(\vec{xy}) = -\theta(\vec{yx})$. For a flow θ, define the ***divergence*** of θ at x by

$$\operatorname{div} \theta(x) := \sum_{y \,:\, y \sim x} \theta(\vec{xy}).$$

We note that for any flow θ we have

$$\sum_{x \in V} \operatorname{div} \theta(x) = \sum_{x \in V} \sum_{y \,:\, y \sim x} \theta(\vec{xy}) = \sum_{\{x,y\} \in E} [\theta(\vec{xy}) + \theta(\vec{yx})] = 0. \tag{9.6}$$

A ***flow from*** a ***to*** z is a flow θ satisfying

(i) ***Kirchhoff's node law***:

$$\operatorname{div} \theta(x) = 0 \quad \text{at all } x \notin \{a, z\}, \tag{9.7}$$

and

(ii) $\operatorname{div} \theta(a) \geq 0$.

Note that (9.7) is the requirement that "flow in equals flow out" for any node not a or z.

We define the ***strength*** of a flow θ from a to z to be $\|\theta\| := \operatorname{div} \theta(a)$. A ***unit flow*** from a to z is a flow from a to z with strength 1. Observe that (9.6) implies that $\operatorname{div} \theta(a) = -\operatorname{div} \theta(z)$.

Observe that it is only flows that are defined on oriented edges. Conductance and resistance are defined for unoriented edges. We may of course define them (for future notational convenience) on oriented edges by $c(\vec{xy}) = c(\vec{yx}) = c(x, y)$ and $r(\vec{xy}) = r(\vec{yx}) = r(x, y)$.

Given a voltage W on the network, the ***current flow*** I associated with W is defined on oriented edges by

$$I(\vec{xy}) = \frac{W(x) - W(y)}{r(\vec{xy})} = c(x, y) \left[W(x) - W(y) \right]. \tag{9.8}$$

Because any voltage is an affine transformation of the unique voltage W_0 satisfying $W_0(a) = 1$ and $W_0(z) = 0$, the unit current flow is unique.

This definition immediately implies that the current flow satisfies ***Ohm's law***:

$$r(\overrightarrow{xy})I(\overrightarrow{xy}) = W(x) - W(y). \tag{9.9}$$

Also notice that I is antisymmetric and satisfies the node law (9.7) at every $x \notin \{a, z\}$:

$$\begin{aligned}\sum_{y\,:\,y\sim x} I(\overrightarrow{xy}) &= \sum_{y\,:\,y\sim x} c(x,y)[W(x) - W(y)] \\ &= c(x)W(x) - c(x) \sum_{y\,:\,y\sim x} W(y)P(x,y) = 0.\end{aligned}$$

Finally, the current flow also satisfies the ***cycle law***. If the oriented edges $\vec{e_1}, \dots, \vec{e_m}$ form an oriented cycle (i.e., for some $x_0, \dots, x_{n-1} \in V$ we have $\vec{e_i} = (x_{i-1}, x_i)$, where $x_n = x_0$), then

$$\sum_{i=1}^{m} r(\vec{e_i})I(\vec{e_i}) = 0\,. \tag{9.10}$$

Notice that adding a constant to all values of a voltage affects neither its harmonicity nor the current flow it determines. Hence we may, without loss of generality, assume our voltage function W satisfies $W(z) = 0$. Such a voltage function is uniquely determined by $W(a)$.

PROPOSITION 9.4 (Node law/cycle law/strength). *If θ is a flow from a to z satisfying the cycle law*

$$\sum_{i=1}^{m} r(\vec{e_i})\theta(\vec{e_i}) = 0 \tag{9.11}$$

for any cycle $\vec{e_1}, \dots, \vec{e_m}$ and if $\|\theta\| = \|I\|$, then $\theta = I$.

PROOF. The function $f = \theta - I$ satisfies the node law at all nodes and the cycle law. Suppose $f(\vec{e_1}) > 0$ for some oriented edge $\vec{e_1}$. By the node law, e_1 must lead to some oriented edge $\vec{e_2}$ with $f(\vec{e_2}) > 0$. Iterate this process to obtain a sequence of oriented edges on which f is strictly positive. Since the underlying network is finite, this sequence must eventually revisit a node. The resulting cycle violates the cycle law. ■

9.4. Effective Resistance

Given a network, the ratio $[W(a) - W(z)]/\|I\|$, where I is the current flow corresponding to the voltage W, is independent of the voltage W applied to the network. Define the ***effective resistance*** between vertices a and z by

$$\mathcal{R}(a \leftrightarrow z) := \frac{W(a) - W(z)}{\|I\|}. \tag{9.12}$$

In parallel with our earlier definitions, we also define the ***effective conductance*** $\mathcal{C}(a \leftrightarrow z) = 1/\mathcal{R}(a \leftrightarrow z)$. Why is $\mathcal{R}(a \leftrightarrow z)$ called the "effective resistance" of the network? Imagine replacing our entire network by a single edge joining a to z with resistance $\mathcal{R}(a \leftrightarrow z)$. If we now apply the same voltage to a and z in both networks, then the amount of current flowing from a to z in the single-edge network is the same as in the original.

Next, we discuss the connection between effective resistance and the ***escape probability*** $\mathbf{P}_a\{\tau_z < \tau_a^+\}$ that a walker started at a hits z before returning to a.

PROPOSITION 9.5. *For any $x, a, z \in \Omega$,*

$$\mathbf{P}_a\{\tau_z < \tau_a^+\} = \frac{1}{c(a)\mathcal{R}(a \leftrightarrow z)} = \frac{\mathcal{C}(a \leftrightarrow z)}{c(a)}. \tag{9.13}$$

PROOF. By Proposition 9.1, the function

$$x \mapsto \mathbf{E}_x \mathbf{1}_{\{X_{\tau_{\{a,z\}}} = z\}} = \mathbf{P}_x\{\tau_z < \tau_a\}$$

is the unique harmonic function on $\Omega \setminus \{a, z\}$ with value 0 at a and value 1 at z. Since the function

$$x \mapsto \frac{W(a) - W(x)}{W(a) - W(z)}$$

is also harmonic on $\Omega \setminus \{a, z\}$ with the same boundary values, we must by Proposition 9.1 have

$$\mathbf{P}_x\{\tau_z < \tau_a\} = \frac{W(a) - W(x)}{W(a) - W(z)}. \tag{9.14}$$

Therefore,

$$\mathbf{P}_a\{\tau_z < \tau_a^+\} = \sum_{x \in V} P(a, x)\mathbf{P}_x\{\tau_z < \tau_a\} = \sum_{x\,:\,x \sim a} \frac{c(a,x)}{c(a)} \frac{W(a) - W(x)}{W(a) - W(z)}. \tag{9.15}$$

By the definition (9.8) of current flow, the above is equal to

$$\frac{\sum_{x\,:\,x \sim a} I(\overrightarrow{ax})}{c(a)\,[W(a) - W(z)]} = \frac{\|I\|}{c(a)\,[W(a) - W(z)]} = \frac{1}{c(a)\mathcal{R}(a \leftrightarrow z)}, \tag{9.16}$$

showing (9.13). ■

The ***Green's function*** for a random walk stopped at a stopping time τ is defined by

$$G_\tau(a, x) := \mathbf{E}_a\,(\text{number of visits to } x \text{ before } \tau) = \mathbf{E}_a\left(\sum_{t=0}^{\infty} \mathbf{1}_{\{X_t = x, \tau > t\}}\right). \tag{9.17}$$

LEMMA 9.6. *If $G_{\tau_z}(a, a)$ is the Green's function defined in* (9.17), *then*

$$G_{\tau_z}(a, a) = c(a)\mathcal{R}(a \leftrightarrow z). \tag{9.18}$$

PROOF. The number of visits to a before visiting z has a geometric distribution with parameter $\mathbf{P}_a\{\tau_z < \tau_a^+\}$. The lemma then follows from (9.13). ■

It is often possible to replace a network by a simplified one without changing quantities of interest, for example the effective resistance between a pair of nodes. The following laws are very useful.

Parallel Law. *Conductances in parallel add*: suppose edges e_1 and e_2, with conductances c_1 and c_2, respectively, share vertices v_1 and v_2 as endpoints. Then both edges can be replaced with a single edge of conductance $c_1 + c_2$ without affecting the rest of the network. All voltages and currents in $G \setminus \{e_1, e_2\}$ are unchanged and the current $I(\overrightarrow{e})$ equals $I(\overrightarrow{e_1}) + I(\overrightarrow{e_2})$. For a proof, check Ohm's and Kirchhoff's laws with $I(\overrightarrow{e}) := I(\overrightarrow{e_1}) + I(\overrightarrow{e_2})$.

Series Law. *Resistances in series add*: if $v \in V \setminus \{a, z\}$ is a node of degree 2 with neighbors v_1 and v_2, the edges (v_1, v) and (v, v_2) can be replaced by a single edge (v_1, v_2) of resistance $r_{v_1 v} + r_{v v_2}$. All potentials and currents in $G \setminus \{v\}$ remain

the same and the current that flows from v_1 to v_2 equals $I(\overrightarrow{v_1 v}) = I(\overrightarrow{v v_2})$. For a proof, check again Ohm's and Kirchhoff's laws, with $I(\overrightarrow{v_1 v_2}) := I(\overrightarrow{v_1 v}) = I(\overrightarrow{v v_2})$.

Gluing. Another convenient operation is to identify vertices having the same voltage, while keeping all existing edges. Because current never flows between vertices with the same voltage, potentials and currents are unchanged.

EXAMPLE 9.7. When a and z are two vertices in a tree Γ with unit resistance on each edge, then $\mathcal{R}(a \leftrightarrow z)$ is equal to the length of the unique path joining a and z. (For any vertex x not along the path joining a and z, there is a unique path from x to a. Let x_0 be the vertex at which the x–a path first hits the a–z path. Then $W(x) = W(x_0)$.)

EXAMPLE 9.8. For a tree Γ with root ρ, let Γ_n be the set of vertices at distance n from ρ. Consider the case of a spherically symmetric tree, in which all vertices of Γ_n have the same degree for all $n \geq 0$. Suppose that all edges at the same distance from the root have the same resistance, that is, $r(e) = r_i$ if $|e| = i$, $i \geq 1$. Glue all the vertices in each level; this will not affect effective resistances, so we infer that

$$\mathcal{R}(\rho \leftrightarrow \Gamma_M) = \sum_{i=1}^{M} \frac{r_i}{|\Gamma_i|} \tag{9.19}$$

and

$$\mathbf{P}_\rho\{\tau_{\Gamma_M} < \tau_\rho^+\} = \frac{r_1/|\Gamma_1|}{\sum_{i=1}^{M} r_i/|\Gamma_i|}. \tag{9.20}$$

Therefore, $\lim_{M\to\infty} \mathbf{P}_\rho\{\tau_{\Gamma_M} < \tau_\rho^+\} > 0$ if and only if $\sum_{i=1}^{\infty} r_i/|\Gamma_i| < \infty$.

EXAMPLE 9.9 (Biased nearest-neighbor random walk). Fix $\alpha > 0$ with $\alpha \neq 1$ and consider the path with vertices $\{0, 1, \ldots, n\}$ and weights $c(k-1, k) = \alpha^k$ for $k = 1, \ldots, n$. Then for all interior vertices $0 < k < n$ we have

$$P(k, k+1) = \frac{\alpha}{1+\alpha},$$
$$P(k, k-1) = \frac{1}{1+\alpha}.$$

If $p = \alpha/(1+\alpha)$, then this is the walk that, when at an interior vertex, moves up with probability p and down with probability $1-p$. (Note: this is also an example of a birth-and-death chain, as defined in Section 2.5.)

Using the Series Law, we can replace the k edges to the left of vertex k by a single edge of resistance

$$r_1 := \sum_{j=1}^{k} \alpha^{-j} = \frac{1-\alpha^{-k}}{\alpha - 1}.$$

Likewise, we can replace the $(n-k)$ edges to the right of k by a single edge of resistance

$$r_2 := \sum_{j=k+1}^{n} \alpha^{-j} = \frac{\alpha^{-k} - \alpha^{-n}}{\alpha - 1}$$

The probability $\mathbf{P}_k\{\tau_n < \tau_0\}$ is not changed by this modification, so we can calculate simply that

$$\mathbf{P}_k\{\tau_n < \tau_0\} = \frac{r_2^{-1}}{r_1^{-1} + r_2^{-1}} = \frac{\alpha^{-k} - 1}{\alpha^{-n} - 1}.$$

In particular, for the biased random walk which moves up with probability p,

$$\mathbf{P}_k\{\tau_n < \tau_0\} = \frac{[(1-p)/p]^k - 1}{[(1-p)/p]^n - 1}. \tag{9.21}$$

Define the ***energy*** of a flow θ by

$$\mathcal{E}(\theta) := \sum_e [\theta(e)]^2 r(e).$$

THEOREM 9.10 (Thomson's Principle). *For any finite connected graph,*

$$\mathcal{R}(a \leftrightarrow z) = \inf\left\{ \mathcal{E}(\theta) : \theta \text{ a unit flow from } a \text{ to } z \right\}. \tag{9.22}$$

The unique minimizer in the inf *above is the unit current flow.*

REMARK 9.11. The sum in $\mathcal{E}(\theta)$ is over unoriented edges, so each edge $\{x, y\}$ is only considered once in the definition of energy. Although θ is defined on oriented edges, it is antisymmetric and hence $\theta(e)^2$ is unambiguous.

PROOF. Since the set of unit-strength flows can be viewed as a closed and bounded subset of $\mathbb{R}^{|E|}$, by compactness there exists a flow θ minimizing $\mathcal{E}(\theta)$ subject to $\|\theta\| = 1$. By Proposition 9.4, to prove that the unit current flow is the unique minimizer, it is enough to verify that any unit flow θ of minimal energy satisfies the cycle law.

Let the edges $\vec{e_1}, \ldots, \vec{e_n}$ form a cycle. Set $\gamma(\vec{e_i}) = 1$ for all $1 \leq i \leq n$ and set γ equal to zero on all other edges. Note that γ satisfies the node law, so it is a flow, but $\sum \gamma(\vec{e_i}) = n \neq 0$. For any $\varepsilon \in \mathbb{R}$, we have by energy minimality that

$$\begin{aligned} 0 \leq \mathcal{E}(\theta + \varepsilon\gamma) - \mathcal{E}(\theta) &= \sum_{i=1}^{n} \left[\left(\theta(\vec{e_i}) + \varepsilon\right)^2 - \theta(\vec{e_i})^2 \right] r(\vec{e_i}) \\ &= 2\varepsilon \sum_{i=1}^{n} r(\vec{e_i})\theta(\vec{e_i}) + O(\varepsilon^2). \end{aligned}$$

Dividing both sides by $\varepsilon > 0$ shows that

$$0 \leq 2\sum_{i=1}^{n} r(\vec{e_i})\theta(\vec{e_i}) + O(\varepsilon),$$

and letting $\varepsilon \downarrow 0$ shows that $0 \leq \sum_{i=1}^{n} r(e_i)\theta(\vec{e_i})$. Similarly, dividing by $\varepsilon < 0$ and then letting $\varepsilon \uparrow 0$ shows that $0 \geq \sum_{i=1}^{n} r(e_i)\theta(\vec{e_i})$. Therefore, $\sum_{i=1}^{n} r(e_i)\theta(\vec{e_i}) = 0$, verifying that θ satisfies the cycle law.

We complete the proof by showing that the unit current flow I has $\mathcal{E}(I) = \mathcal{R}(a \leftrightarrow z)$:

$$\begin{aligned} \sum_e r(e)I(e)^2 &= \frac{1}{2}\sum_x \sum_y r(x,y) \left[\frac{W(x) - W(y)}{r(x,y)}\right]^2 \\ &= \frac{1}{2}\sum_x \sum_y c(x,y)[W(x) - W(y)]^2 \\ &= \frac{1}{2}\sum_x \sum_y [W(x) - W(y)] I(\vec{xy}). \end{aligned}$$

Since I is antisymmetric,

$$\frac{1}{2}\sum_x\sum_y [W(x)-W(y)]I(\overrightarrow{xy}) = \sum_x W(x)\sum_y I(\overrightarrow{xy}). \tag{9.23}$$

By the node law, $\sum_y I(\overrightarrow{xy}) = 0$ for any $x \notin \{a,z\}$, while $\sum_y I(\overrightarrow{ay}) = \|I\| = -\sum_y I(\overrightarrow{zy})$, so the right-hand side of (9.23) equals

$$\|I\|\,(W(a)-W(z)).$$

Since $\|I\| = 1$, we conclude that the right-hand side of (9.23) is equal to $(W(a)-W(z))/\|I\| = \mathcal{R}(a\leftrightarrow z)$. ■

Let a, z be vertices in a network and suppose that we add to the network an edge which is not incident to a. How does this affect the escape probability from a to z? From the point of view of probability, the answer is not obvious. In the language of electrical networks, this question is answered by Rayleigh's Law.

If $r = \{r(e)\}$ is a set of resistances on the edges of a graph G, write $\mathcal{R}(a\leftrightarrow z; r)$ to denote the effective resistance computed with these resistances.

THEOREM 9.12 (Rayleigh's Monotonicity Law). *If $\{r(e)\}$ and $\{r'(e)\}$ are sets of resistances on the edges of the same graph G and if $r(e) \le r'(e)$ for all e, then*

$$\mathcal{R}(a\leftrightarrow z; r) \;\le\; \mathcal{R}(a\leftrightarrow z; r'). \tag{9.24}$$

PROOF. Note that $\inf_\theta \sum_e r(e)\theta(e)^2 \le \inf_\theta \sum_e r'(e)\theta(e)^2$ and apply Thomson's Principle (Theorem 9.10). ■

COROLLARY 9.13. *Adding an edge does not increase the effective resistance $\mathcal{R}(a\leftrightarrow z)$. If the added edge is not incident to a, the addition does not decrease the escape probability $\mathbf{P}_a\{\tau_z < \tau_a^+\} = [c(a)\mathcal{R}(a\leftrightarrow z)]^{-1}$.*

PROOF. Before we add an edge to a network, we can think of it as existing already with $c = 0$ or $r = \infty$. By adding the edge, we reduce its resistance to a finite number.

Combining this with the relationship (9.13) shows that the addition of an edge not incident to a (which we regard as changing a conductance from 0 to a non-zero value) cannot decrease the escape probability $\mathbf{P}_a\{\tau_z < \tau_a^+\}$. ■

COROLLARY 9.14. *The operation of gluing vertices cannot increase effective resistance.*

PROOF. When we glue vertices together, we take an infimum in Thomson's Principle (Theorem 9.10) over a larger class of flows. ■

A technique due to Nash-Williams often gives simple but useful lower bounds on effective resistance. We call $\Pi \subseteq V$ an ***edge-cutset separating** a **from** z* if every path from a to z includes some edge in Π.

PROPOSITION 9.15. *If $\{\Pi_k\}$ are disjoint edge-cutsets which separate nodes a and z, then*

$$\mathcal{R}(a\leftrightarrow z) \ge \sum_k \left(\sum_{e\in\Pi_k} c(e)\right)^{-1}. \tag{9.25}$$

The inequality (9.25) *is called the* Nash-Williams inequality.

PROOF. Let θ be a unit flow from a to z. For any k, by the Cauchy-Schwarz inequality

$$\sum_{e\in\Pi_k} c(e) \cdot \sum_{e\in\Pi_k} r(e)\theta(e)^2 \geq \left(\sum_{e\in\Pi_k} \sqrt{c(e)}\sqrt{r(e)}|\theta(e)|\right)^2 = \left(\sum_{e\in\Pi_k} |\theta(e)|\right)^2.$$

The right-hand side is bounded below by $\|\theta\|^2 = 1$, because Π_k is a cutset and $\|\theta\| = 1$. Therefore

$$\sum_e r(e)\theta(e)^2 \geq \sum_k \sum_{e\in\Pi_k} r(e)\theta(e)^2 \geq \sum_k \left(\sum_{e\in\Pi_k} c(e)\right)^{-1}.$$

By Thompson's Principle (Theorem 9.10), we are done. ∎

9.5. Escape Probabilities on a Square

We now use the inequalities we have developed to bound effective resistance in a non-trivial example. Let B_n be the $n \times n$ two-dimensional grid graph: the vertices are pairs of integers (z, w) such that $1 \leq z, w \leq n$, while the edges are pairs of points at unit (Euclidean) distance.

PROPOSITION 9.16. *Let $a = (1,1)$ be the lower left-hand corner of B_n, and let $z = (n,n)$ be the upper right-hand corner of B_n. Suppose each edge of B_n has unit conductance. The effective resistance $\mathcal{R}(a \leftrightarrow z)$ satisfies*

$$\frac{\log(n-1)}{2} \leq \mathcal{R}(a \leftrightarrow z) \leq 2\log n. \tag{9.26}$$

We separate the proof into the lower and upper bounds.

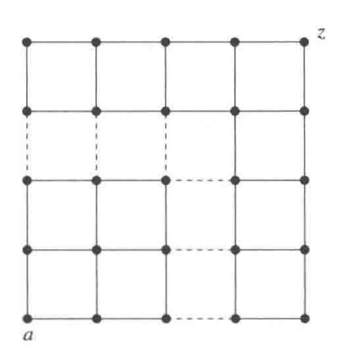


FIGURE 9.1. The graph B_5. The cutset Π_3 contains the edges drawn with dashed lines.

PROOF OF LOWER BOUND IN (9.26). Let Π_k be the edge set

$$\Pi_k = \{(v,w) \in B_n : \|v\|_\infty = k-1, \|w\|_\infty = k\},$$

where $\|(v_1, v_2)\|_\infty = \max\{v_1, v_2\}$. (See Figure 9.1.) Since every path from a to z must use an edge in Π_k, the set Π_k is a cutset. Since each edge has unit conductance,

$\sum_{e \in \Pi_k} c(e)$ equals the number of edges in Π_k, namely $2k$. By Proposition 9.15 and Exercise 2.4,

$$\mathcal{R}(a \leftrightarrow z) \geq \sum_{k=1}^{n-1} \frac{1}{2k} \geq \frac{\log(n-1)}{2}. \tag{9.27}$$

■

PROOF OF UPPER BOUND IN (9.26). Thomson's Principle (Theorem 9.10) says that the effective resistance is the minimal possible energy of a unit flow from a to z. So to get an upper bound on resistance, we build a unit flow on the square.

Consider Pólya's urn process, described in Section 2.4. The sequence of ordered pairs listing the numbers of black and white balls is a Markov chain with state space $\{1, 2, \ldots\}^2$.

Run this process on the square—note that it necessarily starts at vertex $a = (1, 1)$—and stop when you reach the main diagonal $x + y = n + 1$. Direct all edges of the square from bottom left to top right and give each edge e on the bottom left half of the square the flow

$$f(e) = \mathbf{P}\{\text{the process went through } e\}.$$

To finish the construction, give the upper right half of the square the symmetrical flow values.

From Lemma 2.6, it follows that for any $k \geq 0$, the Pólya's urn process is equally likely to pass through each of the $k+1$ pairs (i, j) for which $i + j = k + 2$. Consequently, when (i, j) is a vertex in the square for which $i + j = k + 2$, the sum of the flows on its incoming edges is $\frac{1}{k+1}$. Thus the energy of the flow f can be bounded by

$$\mathcal{E}(f) \leq \sum_{k=1}^{n-1} 2 \left(\frac{1}{k+1} \right)^2 (k+1) \leq 2 \log n.$$

■

Exercises

EXERCISE 9.1. Generalize the flow in the upper bound of (9.26) to higher dimensions, using an urn with balls of d colors. Use this to show that the resistance between opposite corners of the d-dimensional box of side length n is bounded independent of n, when $d \geq 3$.

EXERCISE 9.2. An Oregon professor has n umbrellas, of which initially $k \in (0, n)$ are at his office and $n - k$ are at his home. Every day, the professor walks to the office in the morning and returns home in the evening. In each trip, he takes an umbrella with him only if it is raining. Assume that in every trip between home and office or back, the chance of rain is $p \in (0, 1)$, independently of other trips.

(a) Asymptotically, in what fraction of his trips does the professor get wet?
(b) Determine the expected number of trips until all n umbrellas are at the same location.
(c) Determine the expected number of trips until the professor gets wet.

EXERCISE 9.3 (Gambler's ruin). In Section 2.1, we defined simple random walk on $\{0, 1, 2, \ldots, n\}$. Use the network reduction laws to show that $\mathbf{P}_x\{\tau_n < \tau_0\} = x/n$.

EXERCISE 9.4. Let θ be a flow from a to z which satisfies both the cycle law and $\|\theta\| = \|I\|$. Define a function h on nodes by

$$h(x) = \sum_{i=1}^{m} [\theta(\vec{e_i}) - I(\vec{e_i})]\, r(\vec{e_i}), \tag{9.28}$$

where $\vec{e_i}, \ldots, \vec{e}_m$ is an arbitrary path from a to x.

(a) Show that h is well-defined and harmonic at all nodes.
(b) Use part (a) to give an alternate proof of Proposition 9.4.

EXERCISE 9.5. Show that if, in a network with source a and sink z, vertices with different voltages are glued together, then the effective resistance from a to z will strictly decrease.

EXERCISE 9.6. Show that $\mathcal{R}(a \leftrightarrow z)$ is a concave function of $\{r(e)\}$.

EXERCISE 9.7. Let B_n be the subset of $\mathbb{Z}^2$ contained in the box of side length $2n$ centered at 0. Let ∂B_n be the set of vertices along the perimeter of the box. Show that

$$\lim_{n\to\infty} \mathbf{P}_0\{\tau_{\partial B_n} < \tau_a^+\} = 0.$$

EXERCISE 9.8. Show that effective resistances form a metric on any network with conductances $\{c(e)\}$.

Hint: The only non-obvious statement is the triangle inequality

$$\mathcal{R}(x \leftrightarrow z) \le \mathcal{R}(x \leftrightarrow y) + \mathcal{R}(y \leftrightarrow z).$$

Adding the unit current flow from x to y to the unit current flow from y to z gives the unit current flow from x to z (check Kirchoff's laws!). Now use the corresponding voltage functions.

Notes

Proposition 9.15 appeared in Nash-Williams (1959).

Further reading. The basic reference for the connection between electrical networks and random walks on graphs is Doyle and Snell (1984), and we borrow here from Peres (1999). For much more, see Lyons and Peres (2008).

For an introduction to (continuous) harmonic functions, see Ahlfors (1978, Chapter 6).

CHAPTER 10

Hitting Times

10.1. Definition

Global maps are often unavailable for real networks that have grown without central organization, such as the Internet. However, sometimes the structure can be queried locally, meaning that given a specific node v, for some cost all nodes connected by a single link to v can be determined. How can such local queries be used to determine whether two nodes v and w can be connected by a path in the network?

Suppose you have limited storage, but you are not concerned about time. In this case, one approach is to start a random walk at v, allow the walk to explore the graph for some time, and observe whether the node w is ever encountered. If the walk visits node w, then clearly v and w must belong to the same connected component of the network. On the other hand, if node w has not been visited by the walk by time t, it is possible that w is not accessible from v—but perhaps the walker was simply unlucky. It is of course important to distinguish between these two possibilities! In particular, when w is connected to v, we desire an estimate of the expected time until the walk visits w starting at v.

Given a Markov chain with state space Ω, it is natural to define the ***hitting time*** τ_A of a subset $A \subseteq \Omega$ to be the first time one of the nodes in A is visited by the chain. In symbols, if (X_t) is the random walk, we set

$$\tau_A := \min\{t \geq 0 : X_t \in A\}.$$

We will simply write τ_w for $\tau_{\{w\}}$, consistent with our notation in Section 1.5.2.

We have already seen the usefulness of hitting times. In Section 1.5.3 we used a variant

$$\tau_x^+ = \min\{t \geq 1 : X_t = x\}$$

(called the ***first return time*** when $X_0 = x$) to build a stationary distribution. In Section 5.3.2, we used the expected absorption time for the random walk on a line segment (computed in Section 2.1) to bound the expected coupling time for the torus. In Section 9.2, we used hitting times to construct harmonic functions satisfying specified boundary conditions.

To connect our discussion of hitting times for random walks on networks to our leitmotif of mixing times, we mention now the problem of estimating the mixing time for two "glued" tori, the graph considered in Example 7.4.

Let V_1 be the collection of nodes in the right-hand torus, and let $v^\star$ be the node connecting the two tori.

When the walk is started at a node x in the left-hand torus, we have

$$\|P^t(x,\cdot)-\pi\|_{\mathrm{TV}} \geq \pi(V_1)-P^t(x,V_1) \geq \frac{1}{2}-\mathbf{P}_x\{X_t \in V_1\} \geq \frac{1}{2}-\mathbf{P}_x\{\tau_{v^\star} \leq t\}. \quad (10.1)$$

If the walk is unlikely to have exited the left-hand torus by time t, then (10.1) shows that $d(t)$ is not much smaller than $1/2$. In view of this, it is not surprising that estimates for $\mathbf{E}_x(\tau_{v^\star})$ are useful for bounding t_{mix} for this chain. These ideas are developed in Section 10.6.

10.2. Random Target Times

LEMMA 10.1 (Random Target Lemma). *For an irreducible Markov chain with state space Ω, transition matrix P, and stationary distribution π, the quantity*

$$\sum_{x\in\Omega} \mathbf{E}_a(\tau_x)\pi(x)$$

does not depend on $a \in \Omega$.

PROOF. For notational convenience, let $h_x(a) = \mathbf{E}_a(\tau_x)$. Observe that if $x \neq a$,

$$h_x(a) = \sum_{y\in\Omega} \mathbf{E}_a(\tau_x \mid X_1 = y)P(a,y) = \sum_{y\in\Omega} (1 + h_x(y))\, P(a,y) = (Ph_x)(a) + 1,$$

so that

$$(Ph_x)(a) = h_x(a) - 1. \tag{10.2}$$

If $x = a$, then

$$\mathbf{E}_a(\tau_a^+) = \sum_{y\in\Omega} \mathbf{E}_a(\tau_a^+ \mid X_1 = y)P(a,y) = \sum_{y\in\Omega} (1 + h_a(y))\, P(a,y) = 1 + (Ph_a)(a).$$

Since $\mathbf{E}_a(\tau_a^+) = \pi(a)^{-1}$,

$$(Ph_a)(a) = \frac{1}{\pi(a)} - 1. \tag{10.3}$$

Thus, letting $h(a) := \sum_{x\in\Omega} h_x(a)\pi(x)$, (10.2) and (10.3) show that

$$(Ph)(a) = \sum_{x\in\Omega} (Ph_x)(a)\pi(x) = \sum_{x\neq a} (h_x(a) - 1)\pi(x) + \pi(a)\left(\frac{1}{\pi(a)} - 1\right).$$

Simplifying the right-hand side and using that $h_a(a) = 0$ yields

$$(Ph)(a) = h(a).$$

That is, h is harmonic. Applying Lemma 1.16 shows that h is a constant function. ■

Consider choosing a state $y \in \Omega$ according to π. Lemma 10.1 says that the expected time to hit the "random target" state y from a specified starting state a does not depend on a. Hence we can define the ***target time*** of an irreducible chain by

$$t_\odot := \sum_{x\in\Omega} \mathbf{E}_a(\tau_x)\pi(x) = \mathbf{E}_\pi(\tau_\pi)$$

(the last version is a slight abuse of our notation for hitting times). Since $t_\odot$ does not depend on the state a, it is true that

$$t_\odot = \sum_{x,y\in\Omega} \pi(x)\pi(y)\mathbf{E}_x(\tau_y) = \mathbf{E}_\pi(\tau_\pi). \tag{10.4}$$

We will often find it useful to estimate the worst-case hitting times between states in a chain. Define

$$t_{\mathrm{hit}} := \max_{x,y\in\Omega} \mathbf{E}_x(\tau_y). \tag{10.5}$$

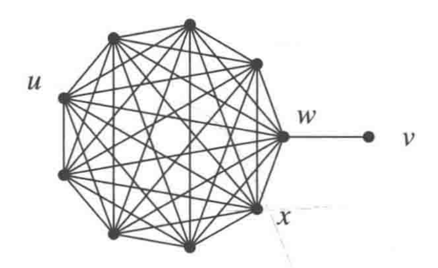


FIGURE 10.1. For random walk on this family of graphs, $t_{\text{hit}} \gg t_{\odot}$.

LEMMA 10.2. *For an irreducible Markov chain with state space* Ω *and stationary distribution* π,

$$t_{\text{hit}} \leq 2 \max_{w} \mathbf{E}_{\pi}(\tau_w).$$

PROOF. For any $a, y \in \Omega$, we have

$$\mathbf{E}_a(\tau_y) \leq \mathbf{E}_a(\tau_\pi) + \mathbf{E}_\pi(\tau_y), \tag{10.6}$$

since we can insist that the chain go from a to y via a random state x chosen according to π. By Lemma 10.1,

$$\mathbf{E}_a(\tau_\pi) = \mathbf{E}_\pi(\tau_\pi) \leq \max_{w} \mathbf{E}_\pi(\tau_w).$$

It is now clear that (10.6) implies the desired inequality. ∎

Note that for a transitive chain, for any b,

$$t_{\odot} = \mathbf{E}_\pi(\tau_\pi) = \sum_{x \in \Omega} \mathbf{E}_a(\tau_x)\pi(x) = \sum_{x,y \in \Omega} \pi(y)\mathbf{E}_y(\tau_x)\pi(x) = \mathbf{E}_\pi(\tau_b).$$

Hence we have

COROLLARY 10.3. *For an irreducible transitive Markov chain,*

$$t_{\text{hit}} \leq 2t_{\odot}.$$

EXAMPLE 10.4. When the underlying chain is not transitive, it is possible for t_{hit} to be much larger than $t_{\odot}$. Consider the example of simple random walk on a complete graph on n vertices with a leaf attached to one vertex (see Figure 10.1). Let v be the leaf and let w be the neighbor of the leaf; call the other vertices ***ordinary***. Let the initial state of the walk be v. The first return time to v satisfies both

$$\mathbf{E}_v\tau_v^+ = \mathbf{E}_v\tau_w + \mathbf{E}_w\tau_v = 1 + \mathbf{E}_w\tau_v$$

(since the walk must take its first step to w) and

$$\mathbf{E}_v\tau_v^+ = \frac{1}{\pi(v)} = \frac{2\binom{n}{2} + 1}{1} = n^2 - n + 2,$$

by Proposition 1.14(ii) and Example 1.12. Hence $\mathbf{E}_w\tau_v = n^2 - n + 1 \leq t_{\text{hit}}$.

By the Random Target Lemma, we can use any starting state to estimate $t_{\odot}$. Let's start at v. Clearly $\mathbf{E}_v\tau_v = 0$, while $\mathbf{E}_v\tau_w = 1$ and $\mathbf{E}_v\tau_u = 1 + \mathbf{E}_w\tau_u$, where u is any ordinary vertex. How long does it take to get from w to u, on average? Let

x be any *other* ordinary vertex. By conditioning on the first step of the walk and exploiting symmetry, we have

$$\begin{aligned}\mathbf{E}_w\tau_u &= 1 + \frac{1}{n}\left(\mathbf{E}_v\tau_u + (n-2)\mathbf{E}_x\tau_u\right)\\ &= 1 + \frac{1}{n}\left(1 + \mathbf{E}_w\tau_u + (n-2)\mathbf{E}_x\tau_u\right)\end{aligned}$$

and

$$\mathbf{E}_x\tau_u = 1 + \frac{1}{n-1}\left(\mathbf{E}_w\tau_u + (n-3)\mathbf{E}_x\tau_u\right).$$

We have two equations in the two "variables" $\mathbf{E}_w\tau_u$ and $\mathbf{E}_x\tau_u$. Solving yields

$$\mathbf{E}_w\tau_u = \frac{n^2-n+4}{n} = O(n) \quad \text{and} \quad \mathbf{E}_x\tau_u = \frac{n^2-n+2}{n} = O(n)$$

(we only care about the first equation right now). Combining these results with Example 1.12 yields

$$\begin{aligned}t_{\odot} = \mathbf{E}_v\tau_\pi &= \pi(v)(0) + \pi(w)(1) + (n-1)\pi(u)O(n)\\ &= \frac{1(0) + n(1) + (n-1)^2 O(n)}{2\left(\binom{n}{2}+1\right)} = O(n) \ll t_{\text{hit}}.\end{aligned}$$

10.3. Commute Time

The ***commute time*** between nodes a and b in a network is the time to move from a to b and then back to a:

$$\tau_{a,b} = \min\{t \geq \tau_b \,:\, X_t = a\}, \tag{10.7}$$

where we assume that $X_0 = a$. The commute time is of intrinsic interest and can be computed or estimated using resistance (the ***commute time identity***, Proposition 10.6). In graphs for which $\mathbf{E}_a(\tau_b) = \mathbf{E}_b(\tau_a)$, the expected hitting time is half the commute time, so estimates for the commute time yield estimates for hitting times. Transitive networks (defined below) enjoy this property (Proposition 10.9).

The following lemma will be used in the proof of the commute time identity:

LEMMA 10.5 (Aldous, Fill). *If τ is a stopping time for a finite and irreducible Markov chain satisfying $\mathbf{P}_a\{X_\tau = a\} = 1$ and $G_\tau(a,x)$ is the Green's function (as defined in* (9.17)*), then*

$$\frac{G_\tau(a,x)}{\mathbf{E}_a(\tau)} = \pi(x) \qquad \textit{for all } x.$$

Exercise 10.1 asks for a proof of Lemma 10.5.

Recall that $\mathcal{R}(a \leftrightarrow b)$ is the effective resistance between the vertices a and b in a network. (Cf. Section 9.4.)

PROPOSITION 10.6 (Commute Time Identity). *Let $(G, \{c(e)\})$ be a network, and let (X_t) be the random walk on this network. For any nodes a and b in V, let $\tau_{a,b}$ be the commute time defined in* (10.7) *between a and b. Then*

$$\mathbf{E}_a(\tau_{a,b}) = \mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_a) = c_G\mathcal{R}(a \leftrightarrow b). \tag{10.8}$$

(Recall that $c(x) = \sum_{y\,:\,y\sim x} c(x,y)$ and that $c_G = \sum_{x\in V} c(x) = 2\sum_{e\in E} c(e)$.)

PROOF. By Lemma 10.5,

$$\frac{G_{\tau_{a,b}}(a,a)}{\mathbf{E}_a(\tau_{a,b})} = \pi(a) = \frac{c(a)}{c_G}.$$

By definition, after visiting b, the chain does not visit a until time $\tau_{a,b}$, so $G_{\tau_{a,b}}(a,a) = G_{\tau_b}(a,a)$. The conclusion follows from Lemma 9.6. ∎

Note that $\mathbf{E}_a(\tau_b)$ and $\mathbf{E}_b(\tau_a)$ can be very different for general Markov chains and even for reversible chains (see Exercise 10.3). However, for certain types of random walks on networks they are equal. A network $\langle G, \{c(e)\}\rangle$ is ***transitive*** if for any pair of vertices $x, y \in V$ there exists a permutation $\psi_{x,y} : V \to V$ with

$$\psi_{x,y}(x) = y \quad \text{and} \quad c(\psi_{x,y}(u), \psi_{x,y}(v)) = c(u,v) \text{ for all } u, v \in V. \tag{10.9}$$

REMARK 10.7. In Section 2.6.2 we defined transitive Markov chains. The reader should check that a random walk on a transitive network is a transitive Markov chain.

Exercise 9.8 shows that the resistances obey a triangle inequality. We can use Proposition 10.6 to provide another proof.

COROLLARY 10.8. *The resistance $\mathcal{R}$ satisfies a triangle inequality: If a, b, c are vertices, then*

$$\mathcal{R}(a \leftrightarrow c) \le \mathcal{R}(a \leftrightarrow b) + \mathcal{R}(b \leftrightarrow c). \tag{10.10}$$

PROPOSITION 10.9. *For a random walk on a transitive connected graph G, for any vertices $a,\ b \in V$,*

$$\mathbf{E}_a(\tau_b) = \mathbf{E}_b(\tau_a). \tag{10.11}$$

Before proving this, it is helpful to establish the following identity:

LEMMA 10.10. *For any three states a, b, c of a reversible Markov chain,*

$$\mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_c) + \mathbf{E}_c(\tau_a) = \mathbf{E}_a(\tau_c) + \mathbf{E}_c(\tau_b) + \mathbf{E}_b(\tau_a).$$

REMARK 10.11. We can reword this lemma as

$$\mathbf{E}_a(\tau_{bca}) = \mathbf{E}_a(\tau_{cba}), \tag{10.12}$$

where τ_{bca} is the time to visit b, then visit c, and then hit a.

A natural approach to proving this is to assume that reversing a sequence started from a and having $\tau_{bca} = n$ yields a sequence started from a having $\tau_{cba} = n$. However, this is not true. For example, the sequence $acabca$ has $\tau_{bca} = 5$, yet the reversed sequence $acbaca$ has $\tau_{cba} = 3$.

PROOF. It turns out that it is much easier to start at stationarity, since it allows us to use reversibility easily. Recall that we use $\mathbf{E}_\pi(\cdot)$ to denote the expectation operator for the chain started with initial distribution π.

Adding $\mathbf{E}_\pi(\tau_a)$ to both sides of (10.12), we find it is enough to show that

$$\mathbf{E}_\pi(\tau_{abca}) = \mathbf{E}_\pi(\tau_{acba}).$$

In fact, we will show equality in distribution, not just expectation. Suppose ξ and $\xi^\star$ are finite strings with letters in V, meaning $\xi \in V^m$ and $\xi^\star \in V^n$ with $m \le n$. We say that $\xi \preceq \xi^\star$ if and only if ξ is a subsequence of $\xi^\star$, that is, there exist indices

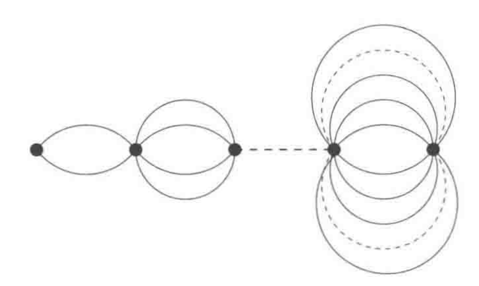


FIGURE 10.2. A binary tree after identifying all vertices at the same distance from the root

$1 \le i_1 < \cdots < i_m \le n$ with $\xi(k) = \xi^\star(i_k)$ for all $1 \le k \le m$. Using the identity (1.32) for reversed chains,

$$\mathbf{P}_\pi\{\tau_{abca} > k\} = \mathbf{P}_\pi\{abca \not\preceq X_0 \ldots X_k\} = \mathbf{P}_\pi\{abca \not\preceq X_k \ldots X_0\}. \tag{10.13}$$

Clearly, $abca \preceq X_k \ldots X_0$ is equivalent to $acba \preceq X_0 \ldots X_k$ (just read from right to left!), so the right-hand side of (10.13) equals

$$\mathbf{P}_\pi\{acba \not\preceq X_0 \ldots X_k\} = \mathbf{P}_\pi\{\tau_{acba} > k\}.$$

■

PROOF OF PROPOSITION 10.9. Let ψ be a map satisfying the conditions (10.9) with $u = a$ and $v = b$. Let $a_0 = a$ and $a_j = \psi^{(j)}(a_0)$ for $j \ge 1$, where $\psi^{(j)}$ denotes the j-th iterate of ψ. The sequence $a_0, a_1, \ldots$ will return to a_0 eventually; say $a_m = a_0$, where $m > 0$. The function $\psi^{(j)}$ takes a, b to a_j, a_{j+1}, so for any j,

$$\mathbf{E}_{a_j}(\tau_{a_{j+1}}) = \mathbf{E}_a(\tau_b). \tag{10.14}$$

Summing over j from 0 to $m-1$, we obtain

$$\mathbf{E}_{a_0}(\tau_{a_1 a_2 \ldots a_{m-1} a_0}) = m\mathbf{E}_a(\tau_b). \tag{10.15}$$

For the same reason,

$$\mathbf{E}_{a_0}(\tau_{a_{m-1} a_{m-2} \ldots a_1 a_0}) = m\mathbf{E}_b(\tau_a). \tag{10.16}$$

By the same argument as we used for (10.12), we see that the left-hand sides of (10.15) and (10.16) are the same. This proves (10.11). ■

EXAMPLE 10.12 (Random walk on rooted finite binary trees). The rooted and finite binary tree of depth k was defined in Section 5.3.4. We let n denote the number of vertices and note that the number of edges equals $n-1$.

We compute the expected commute time between the root and the set of leaves B. Identify all vertices at level j for $j = 1$ to k to obtain the graph shown in Figure 10.2.

Using the network reduction rules, this is equivalent to a segment of length k, with conductance between $j-1$ and j equal to 2^j for $1 \le j \le k$. Thus the effective resistance from the root to the set of leaves B equals

$$\mathcal{R}(a \leftrightarrow B) = \sum_{j=1}^{k} 2^{-j} = 1 - 2^{-k} \le 1.$$

Using the Commute Time Identity (Proposition 10.6), since $c_G = 2(n-1)$, the expected commute time is bounded by $2n$. For the lazy random walk, the expected commute time is bounded by $4n$.

This completes the proof in Section 5.3.4 that $t_{\text{mix}} \leq 16n$.

10.4. Hitting Times for the Torus

Since the torus is transitive, Proposition 10.9 and the Commute Time Identity (Proposition 10.6) imply that for random walk on the d-dimensional torus,

$$\mathbf{E}_a(\tau_b) = 2n^d \mathcal{R}(a \leftrightarrow b). \tag{10.17}$$

(For an unweighted graph, $c = 2 \times |\text{edges}|$.) Thus, to get estimates on the hitting time $\mathbf{E}_a(\tau_b)$, it is enough to get estimates on the effective resistance.

PROPOSITION 10.13. *Let x and y be two points at distance $k \geq 1$ in the torus $\mathbb{Z}_n^d$, and let τ_y be the time of the first visit to y. There exist constants $0 < c_d \leq C_d < \infty$ such that*

$$c_d n^d \leq \mathbf{E}_x(\tau_y) \leq C_d n^d \qquad \textit{uniformly in } k \textit{ if } d \geq 3, \tag{10.18}$$

$$c_2 n^2 \log(k) \leq \mathbf{E}_x(\tau_y) \leq C_2 n^2 \log(k+1) \qquad \textit{if } d = 2. \tag{10.19}$$

PROOF. First, the lower bounds. Choose Π_j to be the boundary of the box centered around x of side-length $2j$. There is a constant $\tilde{c}_1$ so that for $j \leq \tilde{c}_1 k$, the box Π_j is a cutset separating x from y. Note that Π_j has order j^{d-1} edges. By Proposition 9.15,

$$\mathcal{R}(a \leftrightarrow z) \geq \sum_{j=1}^{\tilde{c}_1 k} \tilde{c}_2 j^{1-d} \geq \begin{cases} \tilde{c}_3 \log(k) & \text{if } d = 2, \\ \tilde{c}_3 & \text{if } d \geq 3. \end{cases}$$

The lower bounds in (10.18) and (10.19) are then immediate from (10.17).

If the points x and y are the diagonally opposite corners of a square, the upper bound in (10.19) follows using the flow constructed from Pólya's urn process, described in Section 2.4. in Proposition 9.16.

Now consider the case where x and y are in the corners of a non-square rectangle, as in Figure 10.3. Suppose that $x = (a, b)$ and $y = (c, d)$, and assume without loss of generality that $a \leq c$, $b \leq d$, $(c - a) \leq (d - b)$. Assume also that $c - a$ and $d - b$ have the same parity. The line with slope -1 through x and the line with slope 1 through y meet at the point

$$z = \left(\frac{(a+c) + (b-d)}{2}, \frac{(a-c) + (b+d)}{2} \right).$$

By Proposition 9.16,

$$\mathcal{R}(y \leftrightarrow z) \leq 2 \log \left(\frac{(c-a) + (d-b)}{2} \right) \leq 2 \log(k+1),$$

$$\mathcal{R}(z \leftrightarrow x) \leq 2 \log \left(\frac{(a-c) + (d-b)}{2} \right) \leq 2 \log(k+1).$$

By the triangle inequality for resistances (Corollary 10.8),

$$\mathcal{R}(x \leftrightarrow y) \leq 4 \log(k+1). \tag{10.20}$$

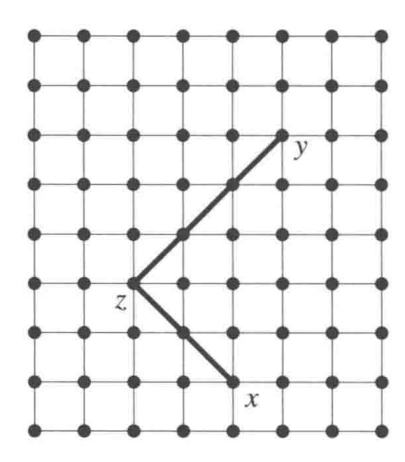


FIGURE 10.3. Constructing a flow from x to y.

When $(c-a)$ and $(d-b)$ have opposite parities, let x' be a lattice point at unit distance from x and closer to y. Applying the triangle inequality again shows that

$$\mathcal{R}(x \leftrightarrow y) \leq \mathcal{R}(x \leftrightarrow x') + \mathcal{R}(x' \leftrightarrow y) \leq 1 + 4\log(k+1) \leq 6\log(k+1). \quad (10.21)$$

Thus (10.20) and (10.21), together with (10.17), establish the upper bound in (10.19). ■

10.5. Bounding Mixing Times via Hitting Times

The goal of this section is to prove the following:

THEOREM 10.14. *Consider a finite reversible chain with transition matrix P and stationary distribution π on Ω.*

(i) *For all $m \geq 0$ and $x \in \Omega$, we have*

$$\|P^m(x,\cdot) - \pi\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left[\frac{P^{2m}(x,x)}{\pi(x)} - 1\right]. \quad (10.22)$$

(ii) *If the chain satisfies $P(x,x) \geq 1/2$ for all x, then*

$$t_{\mathrm{mix}}(1/4) \leq 2\max_{x\in\Omega} \mathbf{E}_\pi(\tau_x) + 1. \quad (10.23)$$

REMARK 10.15. Part (i) says that the total variation distance to stationarity starting from x, for reversible chains, can be made small just by making the return time to x close to its stationary probability.

REMARK 10.16. Note that by conditioning on X_0,

$$\mathbf{E}_\pi(\tau_x) = \sum_{y\in\Omega} \mathbf{E}_y(\tau_x)\pi(y) \leq \max_{y\in\Omega} \mathbf{E}_y(\tau_x) \leq t_{\mathrm{hit}}.$$

Thus the bound (10.23) implies

$$t_{\mathrm{mix}}(1/4) \leq 2t_{\mathrm{hit}} + 1. \quad (10.24)$$

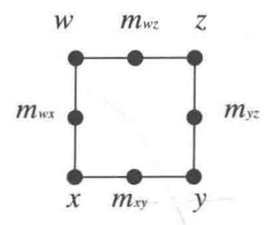


FIGURE 10.4. Adding states m_{xy} for each pair $x, y \in \Omega$.

REMARK 10.17. Equation 10.23 may not hold if the chain is not reversible; see Exercise 10.15. However, a similar inequality for the Cesaro mixing time $t^{\star}_{\text{mix}}$ (defined in Section 6.6) does not require laziness or reversibility: Theorem 6.15 implies that

$$t^{\star}_{\text{mix}}(1/4) \leq 4t_{\text{hit}} + 1$$

for any irreducible chain.

To prove Theorem 10.14, we will need a few preliminary results.

PROPOSITION 10.18. *Let P be the transition matrix for a finite reversible chain on state space Ω with stationary distribution π.*

(i) *For all $t \geq 0$ and $x \in \Omega$ we have $P^{2t+2}(x,x) \leq P^{2t}(x,x)$.*
(ii) *If the chain P_L is lazy, that is $P_L(x,x) \geq 1/2$ for all x, then for all $t \geq 0$ and $x \in \Omega$ we have $P_L^{t+1}(x,x) \leq P_L^t(x,x)$.*

See Exercise 12.6 for a proof using eigenvalues. Here, we give a direct proof using the Cauchy-Schwarz inequality.

PROOF. (i) Since $P^{2t+2}(x,x) = \sum_{y,z\in\Omega} P^t(x,y)P^2(y,z)P^t(z,x)$, we have

$$\pi(x)P^{2t+2}(x,x) = \sum_{y,z\in\Omega} P^t(y,x)\pi(y)P^2(y,z)P^t(z,x) = \sum_{y,z\in\Omega} \psi(y,z)\psi(z,y), \tag{10.25}$$

where $\psi(y,z) = P^t(y,x)\sqrt{\pi(y)P^2(y,z)}$. (By Exercise 1.8, the matrix P^2 is reversible with respect to π.)

By Cauchy-Schwarz, the right-hand side of (10.25) is at most

$$\sum_{y,z\in\Omega} \psi(y,z)^2 = \sum_{y\in\Omega} [P^t(y,x)]^2\pi(y) = \pi(x)P^{2t}(x,x).$$

(ii) Given a lazy chain $P_L = (P+I)/2$, enlarge the state space by adding a new state $m_{xy} = m_{yx}$ for each pair of states $x, y \in \Omega$. (See Figure 10.4.)

On the larger state space Ω_K define a transition matrix K by

$$\begin{aligned} K(x, m_{xy}) &= P(x,y) && \text{for } x, y \in \Omega, \\ K(m_{xy}, x) = K(m_{xy}, y) &= 1/2 && \text{for } x \neq y, \\ K(m_{xx}, x) &= 1 && \text{for all } x, \end{aligned}$$

other transitions having K-probability 0. Then K is reversible with stationary measure π_K given by $\pi_K(x) = \pi(x)/2$ for $x \in \Omega$ and $\pi_K(m_{xy}) = \pi(x)P(x,y)$. Clearly $K^2(x,y) = P_L(x,y)$ for $x, y \in \Omega$, so $K^{2t}(x,y) = P_L^t(x,y)$, and the claimed monotonicity follows. ∎

The following proposition, which does not require reversibility, relates the mean hitting time of a state x to return probabilities.

PROPOSITION 10.19 (Hitting time from stationarity). *Consider a finite irreducible aperiodic chain with transition matrix P with stationary distribution π on Ω. Then for any $x \in \Omega$,*

$$\pi(x)\mathbf{E}_\pi(\tau_x) = \sum_{t=0}^{\infty}[P^t(x,x) - \pi(x)]. \tag{10.26}$$

We give two proofs, one using generating functions and one using stopping times, following (Aldous and Fill, 1999, Lemma 11, Chapter 2).

PROOF OF PROPOSITION 10.19 VIA GENERATING FUNCTIONS. Define

$$f_k := \mathbf{P}_\pi\{\tau_x = k\} \quad \text{and} \quad u_k := P^k(x,x) - \pi(x).$$

Since $\mathbf{P}_\pi\{\tau_x = k\} \leq \mathbf{P}_\pi\{\tau_x \geq k\} \leq C\alpha^k$ for some $\alpha < 1$ (see (1.18)), the power series $F(s) := \sum_{k=0}^\infty f_k s^k$ converges in an interval $[0, 1+\delta_1]$ for some $\delta_1 > 0$.

Also, since $|P^k(x,x) - \pi(x)| \leq d(k)$ and $d(k)$ decays at least geometrically fast (Theorem 4.9), $U(s) := \sum_{k=0}^\infty u_k s^k$ converges in an interval $[0, 1+\delta_2]$ for some $\delta_2 > 0$. Note that $F'(1) = \sum_{k=0}^\infty k f_k = \mathbf{E}_\pi(\tau_x)$ and $U(1)$ equals the right-hand side of (10.26).

For every $m \geq 0$,

$$\pi(x) = \mathbf{P}_\pi\{X_m = x\} = \sum_{k=0}^{m} f_k P^{m-k}(x,x) = \sum_{k=0}^{m} f_k\left[\left(P^{m-k}(x,x) - \pi(x)\right) + \pi(x)\right]$$
$$= \sum_{k=0}^{m} f_k[u_{m-k} + \pi(x)].$$

Thus, the constant sequence with every element equal to $\pi(x)$ is the convolution of the sequence $\{f_k\}_{k=0}^\infty$ with the sequence $\{u_k + \pi(x)\}_{k=0}^\infty$, so its generating function $\sum_{m=0}^\infty \pi(x)s^m = \pi(x)(1-s)^{-1}$ equals the product of the generating function F with the generating function

$$\sum_{m=0}^{\infty}[u_m + \pi(x)]s^m = U(s) + \pi(x)\sum_{m=0}^{\infty} s^m = U(S) + \frac{\pi(x)}{1-s}.$$

(See Exercise 10.9.) That is, for $0 < s < 1$,

$$\frac{\pi(x)}{1-s} = \sum_{m=0}^{\infty}\pi(x)s^m = F(s)\left[U(s) + \frac{\pi(x)}{1-s}\right],$$

and multiplying by $1-s$ gives $\pi(x) = F(s)[(1-s)U(s) + \pi(x)]$, which clearly holds also for $s = 1$. Differentiating the last equation at $s = 1$, we obtain that $0 = F'(1)\pi(x) - U(1)$, and this is equivalent to (10.26). ■

PROOF OF PROPOSITION 10.19 VIA STOPPING TIMES. Define

$$\tau_x^{(m)} := \min\{t \geq m : X_t = x\},$$

and write $\mu_m := P^m(x,\cdot)$. By the Convergence Theorem (Theorem 4.9), μ_m tends to π as $m \to \infty$. By Lemma 10.5, we can represent the expected number of visits

to x before time $\tau_x^{(m)}$ as

$$\sum_{k=0}^{m-1} P^k(x,x) = \pi(x)\mathbf{E}_x\left(\tau_x^{(m)}\right) = \pi(x)[m + \mathbf{E}_{\mu_m}(\tau_x)].$$

Thus $\sum_{k=0}^{m-1}[P^k(x,x) - \pi(x)] = \pi(x)\mathbf{E}_{\mu_m}(\tau_x)$.

Taking $m \to \infty$ completes the proof. ∎

We are now able to prove Theorem 10.14.

PROOF OF THEOREM 10.14. (i) By Cauchy-Schwarz,

$$\left(\frac{1}{2}\sum_{y\in\Omega}\pi(y)\left|\frac{P^m(x,y)}{\pi(y)} - 1\right|\right)^2 \le \frac{1}{4}\sum_{y\in\Omega}\pi(y)\left[\frac{P^m(x,y)}{\pi(y)} - 1\right]^2.$$

Therefore

$$\|P^m(x,\cdot) - \pi\|_{\mathrm{TV}}^2 \le \frac{1}{4}\sum_{y\in\Omega}\left[\frac{P^m(x,y)P^m(y,x)}{\pi(x)} - 2P^m(x,y) + 1\right]$$
$$= \frac{1}{4}\left[\frac{P^{2m}(x,x)}{\pi(x)} - 1\right].$$

(ii) By the identity (10.26) in Proposition 10.19 and the monotonicity in Proposition 10.18(ii), for any $m > 0$ we have

$$\pi(x)\mathbf{E}_\pi(\tau_x) \ge \sum_{k=1}^{2m}[P^k(x,x) - \pi(x)] \ge 2m[P^{2m}(x,x) - \pi(x)].$$

Dividing by $8m\,\pi(x)$ and invoking (10.22) gives

$$\frac{\mathbf{E}_\pi(\tau_x)}{8m} \ge \|P^m(x,\cdot) - \pi\|_{\mathrm{TV}}^2,$$

and the left-hand side is less than $1/16$ for $m \ge 2\mathbf{E}_\pi(\tau_x)$. ∎

EXAMPLE 10.20 (Lazy random walk on the cycle). In Section 5.3.1 we proved that $t_{\mathrm{mix}} \le n^2$ for the lazy random walk on the cycle $\mathbb{Z}_n$. However, Theorem 10.14 can also be used.

Label the states of $\mathbb{Z}_n$ with $\{0, 1, \ldots, n-1\}$. By identifying the states 0 and n, we can see that $\mathbf{E}_k(\tau_0)$ for the lazy simple random walk on the cycle must be the same as the expected time to ruin or success in a lazy gambler's ruin on the path $\{0, 1, \ldots, n\}$. Hence, for lazy simple random walk on the cycle, Exercise 2.1 implies

$$t_{\mathrm{hit}} = \max_{x,y}\mathbf{E}_x(\tau_y) = \max_{0\le k\le n} 2k(n-k) = \left\lfloor \frac{n^2}{2} \right\rfloor.$$

(The factor of 2 comes from the laziness.) Therefore, (10.24) gives

$$t_{\mathrm{mix}} \le n^2 + 1.$$

10.6. Mixing for the Walk on Two Glued Graphs

For a graph $G = (V, E)$ and a vertex $v_\star \in V$, define the set

$$W = \{(v, i) : v \in V,\ i \in \{1, 2\}\}, \tag{10.27}$$

with the elements $(v_\star, 1)$ and $(v_\star, 2)$ identified. Let H be the graph with vertex set W and edge set

$$\{\{(v, i), (w, j)\} : \{v, w\} \in E,\ i = j\}. \tag{10.28}$$

Think of H as two copies of G joined together at the single vertex $v_\star$.

We state the main result of this section:

Proposition 10.21. *For a graph G, let H be the graph with the vertex set W defined in* (10.27) *and edge set defined in* (10.28). *Let τ^G_{couple} be the time for a coupling of two random walks on G to meet. Then there is a coupling of two random walks on H which has a coupling time τ^H_{couple} satisfying*

$$\max_{u,v \in H} \mathbf{E}_{u,v}\left(\tau^H_{\text{couple}}\right) \le \max_{x,y \in G} \mathbf{E}\left(\tau^G_{\text{couple}}\right) + \max_{x \in G} \mathbf{E}_x\left(\tau^G_{v_\star}\right). \tag{10.29}$$

(Here $\tau^G_{v_\star}$ is the hitting time of $v_\star$ in the graph G.)

Proof. Let $\psi : W \to V$ be defined by $\psi(v, i) = v$, and let $\varphi : W \to \{1, 2\}$ be defined by $\varphi(v, i) = i$.

Given a random walk (X^0_t) on G, we show now that a random walk (X_t) can be defined on H satisfying $\psi(X_t) = X^0_t$. Suppose that $(X_s)_{s \le t}$ has already been defined and that $\varphi(X_t) = i$. If $X^0_t \ne v_\star$, then define $X_{t+1} = (X^0_{t+1}, i)$. If $X^0_t = v_\star$, then toss a coin, and define

$$X_{t+1} = \begin{cases} (X^0_{t+1}, 1) & \text{if heads,} \\ (X^0_{t+1}, 2) & \text{if tails.} \end{cases}$$

We now define a coupling (X_t, Y_t) of two random walks on H. Let (X^0_t, Y^0_t) be a coupling of two random walks on G. Until time

$$\tau^G_{\text{couple}} := \min\{t \ge 0 : X^0_t = Y^0_t\},$$

define $(X_t)_{t \le \tau^G_{\text{couple}}}$ and $(Y_t)_{t \le \tau^G_{\text{couple}}}$ by lifting the walks (X^0_t) and (Y^0_t) to H via the procedure described above.

If $X_{\tau^G_{\text{couple}}} = Y_{\tau^G_{\text{couple}}}$, then let (X_t) and (Y_t) evolve together for $t \ge \tau^G_{\text{couple}}$.

Suppose, without loss of generality, that $\varphi(X_{\tau^G_{\text{couple}}}) = 1$ and $\varphi(Y_{\tau^G_{\text{couple}}}) = 2$. Until time

$$\tau_{(v^\star,1)} := \inf\{t \ge \tau^G_{\text{couple}} : X_t = (v_\star, 1)\},$$

couple (Y_t) to (X_t) by setting $Y_t = (\psi(X_t), 2)$. Observe that $\tau_{(v^\star,1)} = \tau^H_{\text{couple}}$, since $(v_\star, 1)$ is identified with $(v_\star, 2)$. The expected difference $\tau^H_{\text{couple}} - \tau^G_{\text{couple}}$ is bounded by $\max_{x \in G} \mathbf{E}_x(\tau_{v_\star})$, whence for $u, v \in H$,

$$\mathbf{E}_{u,v}(\tau^H_{\text{couple}}) \le \mathbf{E}_{\psi(u),\psi(v)}(\tau^G_{\text{couple}}) + \max_{x \in G} \mathbf{E}_x(\tau_{v_\star}).$$

■

We can now return to the example mentioned in this chapter's introduction:

COROLLARY 10.22. *Consider the lazy random walk on two tori glued at a single vertex. (See Example 7.4 and in particular Figure 7.2.) There are constants c_1, c_2 such that*

$$c_1 n^2 \log n \leq t_{\text{mix}} \leq c_2 n^2 \log n, \tag{10.30}$$

where t_{mix} is the mixing time defined in (4.33).

PROOF OF UPPER BOUND IN (10.30). Applying Proposition 10.21, using the bounds in Proposition 10.13 and the bound (5.8) for the coupling on the torus used in Theorem 5.5 shows that there is a coupling with

$$\max_{x,y \in G} \mathbf{E}_{x,y}(\tau_{\text{couple}}) \leq C_1 n^2 \log n. \tag{10.31}$$

Applying Theorem 5.2 shows that

$$\bar{d}(t) \leq \frac{C_1 n^2 \log n}{t},$$

proving the right-hand inequality in (10.30). ■

Exercises

EXERCISE 10.1. Prove Lemma 10.5 by copying the proof in Proposition 1.14 that $\tilde{\pi}$ as defined in (1.19) satisfies $\tilde{\pi} = \tilde{\pi} P$, substituting $G_\tau(a, x)$ in place of $\tilde{\pi}(x)$.

EXERCISE 10.2. Consider the problem of waiting for the sequence TTT to appear in a sequence of fair coin tosses. Is this the same as the waiting time for the sequence HTH?

These waiting times are hitting times for a Markov chain: let X_t be the triplet consisting of the outcomes of tosses $(t, t+1, t+2)$. Then (X_t) is a Markov chain, and the waiting time for TTT is a hitting time. Find $\mathbf{E}(\tau_{TTT})$ and $\mathbf{E}(\tau_{HTH})$.

EXERCISE 10.3. Let G be a connected graph on at least 3 vertices in which the vertex v has only one neighbor, namely w. Show that for the simple random walk on G, $\mathbf{E}_v \tau_w \neq \mathbf{E}_w \tau_v$.

EXERCISE 10.4. Consider simple random walk on the binary tree of depth k with $n = 2^{k+1} - 1$ vertices (first defined in Section 5.3.4).

(a) Let a and b be two vertices at level m whose most recent common ancestor c is at level $h < m$. Show that $\mathbf{E}_a \tau_b = \mathbf{E} \tau_{a,c}$, and find its value.
(b) Show that the maximal value of $\mathbf{E}_a \tau_b$ is achieved when a and b are leaves whose most recent common ancestor is the root of the tree.

EXERCISE 10.5. Let $\mathbf{0} = (0, 0, \ldots, 0)$ be the all-zero vector in the m-dimensional hypercube $\{0,1\}^m$, and let v be a vertex with Hamming weight k. Write $h_m(k)$ for the expected hitting time from v to $\mathbf{0}$ for simple (that is, not lazy) random walk on the hypercube. Determine $h_m(1)$ and $h_m(m)$. Deduce that both $\min_{k>0} h_m(k)$ and $\max_{k>0} h_m(k)$ are asymptotic to 2^m as m tends to infinity. (We say that $f(m)$ is asymptotic to $g(m)$ if their ratio tends to 1.)

Hint: Consider the multigraph G_m obtained by gluing together all vertices of Hamming weight k for each k between 1 and $m-1$. This is a graph on the vertex set $\{0, 1, 2, \ldots, m\}$ with $k\binom{m}{k}$ edges from $k-1$ to k.

EXERCISE 10.6. Use Proposition 10.21 to bound the mixing time for two hypercubes identified at a single vertex.

EXERCISE 10.7. Let (X_t) be a random walk on a network with conductances $\{c_e\}$. Show that

$$\mathbf{E}_a(\tau_{bca}) = [\mathcal{R}(a \leftrightarrow b) + \mathcal{R}(b \leftrightarrow c) + \mathcal{R}(c \leftrightarrow a)] \sum_{e \in E} c_e,$$

where τ_{bca} is the first time that the sequence (b, c, a) appears as a subsequence of $(X_1, X_2, \ldots)$.

EXERCISE 10.8. Show that for a random walk (X_t) on a network, for every three vertices a, x, z,

$$\mathbf{P}_x\{\tau_z < \tau_a\} = \frac{\mathcal{R}(a \leftrightarrow x) - \mathcal{R}(x \leftrightarrow z) + \mathcal{R}(a \leftrightarrow z)}{2\mathcal{R}(a \leftrightarrow z)}.$$

Hint: Run the chain from x until it first visits a and then z. This will also be the first visit to z from x, unless $\tau_z < \tau_a$. In the latter case the path from x to a to z involves an extra commute from z to a beyond time τ_z. Thus, starting from x we have

$$\tau_{az} = \tau_z + \mathbf{1}_{\{\tau_z < \tau_a\}} \tau'_{az}, \tag{10.32}$$

where the variable τ'_{az} refers to the chain starting from its first visit to z. Now take expectations and use the cycle identity (Lemma 10.10).

EXERCISE 10.9. Suppose that $\{a_k\}$ is a sequence with generating function $A(s) := \sum_{k=0}^{\infty} a_k s^k$ and $\{b_k\}$ is a sequence with generating function $B(s) := \sum_{k=0}^{\infty} b_k s^l$. Let $\{c_k\}$ be the sequence defined as $c_k := \sum_{j=0}^{k} a_j b_{k-j}$, called the ***convolution*** of $\{a_k\}$ and $\{b_k\}$. Show that the generating function of $\{c_k\}$ equals $A(s)B(s)$.

EXERCISE 10.10. Let $\tau_x^{\sharp}$ denote the first even time that the Markov chain visits x. Prove that the inequality

$$t_{\text{mix}}(1/4) \le 8 \max_{x \in \Omega} \mathbf{E}_\pi\left(\tau_x^{\sharp}\right) + 1$$

holds without assuming the chain is lazy (cf. Theorem 10.14).

EXERCISE 10.11. Show that for simple random walk (not necessarily lazy) on $\mathbb{Z}_n$, with n odd, $t_{\text{mix}} = O(n^2)$.

Hint: Use Exercise 10.10.

EXERCISE 10.12. Recall the Cesaro mixing time $t_{\text{mix}}^{\star}$ defined in Section 6.6. Show that $t_{\text{mix}}^{\star}(1/4) \le 6 t_{\text{mix}}(1/8)$.

EXERCISE 10.13. Show that $t_{\text{mix}}^{\star}(2^{-k}) \le k t_{\text{mix}}^{\star}(1/4)$ for all $k \ge 1$.

EXERCISE 10.14. Consider a lazy biased random walk on the n-cycle. That is, at each time $t \ge 1$, the particle walks one step clockwise with probability $p \in (1/4, 1/2)$, stays put with probability $1/2$, and walks one step counterclockwise with probability $1/2 - p$.

Show that $t_{\text{mix}}(1/4)$ is bounded above and below by constant multiples of n^2, but $t_{\text{mix}}^{\star}(1/4)$ is bounded above and below by constant multiples of n.

EXERCISE 10.15. Show that equation (10.23) may not hold if the chain is not reversible.

Hint: Consider the lazy biased random walk on the cycle.

EXERCISE 10.16. Suppose that τ is a strong stationary time for simple random walk started at the vertex v on the graph G. Let H consist of two copies G_1 and G_2 of G, glued at v. Note that $\deg_H(v) = 2\deg_G(v)$. Let τ_v be the hitting time of v:

$$\tau_v = \min\{t \geq 0 \, : \, X_t = v\}.$$

Show that starting from any vertex x in H, the random time $\tau_v + \tau$ is a strong stationary time for H (where τ is applied to the walk after it hits v).

REMARK 10.23. It is also instructive to give a general direct argument controlling mixing time in the graph H described in Exercise 10.16:

Let $h_{\max}$ be the maximum expected hitting time of v in G, maximized over starting vertices. For $t > 2kh_{\max} + t_{\mathrm{mix}G}(\varepsilon)$ we have in H that

$$|P^t(x, A) - \pi(A)| < 2^{-k} + \varepsilon. \tag{10.33}$$

Indeed for all x in H, we have $\mathbf{P}_x\{\tau_v > 2h_{\max}\} < 1/2$ and iterating, $\mathbf{P}_x\{\tau_v > 2kh_{\max}\} < (1/2)^k$. On the other hand, conditioning on $\tau_v < 2kh_{\max}$, the bound (10.33) follows from considering the projected walk.

Notes

For more on waiting times for patterns in coin tossing, see Section 17.3.2.

The mean commute identity appears in Chandra, Raghavan, Ruzzo, Smolensky, and Tiwari (1996/97).

Theorem 10.14 is a simplified version of Lemma 15 in Aldous and Fill (1999, Chapter 4), which bounds t_{mix} by $O(t_{\odot})$.

A graph similar to our glued tori was analyzed in Saloff-Coste (1997, Section 3.2) using other methods. This graph originated in Diaconis and Saloff-Coste (1996, Remark 6.1).

CHAPTER 11

Cover Times

11.1. Cover Times

Let (X_t) be a finite Markov chain with state space Ω. The ***cover time*** τ_{cov} of (X_t) is the first time at which all the states have been visited. More formally, τ_{cov} is the minimal value such that, for every state $y \in \Omega$, there exists $t \leq \tau_{\text{cov}}$ with $X_t = y$.

We also define a deterministic version of the cover time by taking the expected value from the worst-case initial state:

$$t_{\text{cov}} = \max_{x \in \Omega} \mathbf{E}_x \tau_{\text{cov}}. \tag{11.1}$$

The cover time of a Markov chain is a natural concept. It can be large enough for relatively small chains to arouse mathematical curiosity. Of course, there are also "practical" interpretations of the cover time. For instance, we might view the progress of a web crawler as a random walk on the graph of World Wide Web pages: at each step, the crawler chooses a linked page at random and goes there. The time taken to scan the entire collection of pages is the cover time of the underlying graph.

EXAMPLE 11.1 (Cover time of cycle). Lovász (1993) gives an elegant computation of the expected cover time t_{cov} of simple random walk on the n-cycle. This walk is simply the remainder modulo n of a simple random walk on $\mathbb{Z}$. The walk on the remainders has covered all n states exactly when the walk on $\mathbb{Z}$ has first visited n distinct states.

Let c_n be the expected value of the time when a simple random walk on $\mathbb{Z}$ has first visited n distinct states, and consider a walk which has just reached its $(n-1)$-st new state. The visited states form a subsegment of the number line and the walk must be at one end of that segment. Reaching the n-th new state is now a gambler's ruin situation: the walker's position corresponds to a fortune of 1 (or $n-1$), and we are waiting for her to reach either 0 or n. Either way, the expected time is $(1)(n-1) = n-1$, as shown in Exercise 2.1. It follows that

$$c_n = c_{n-1} + (n-1) \quad \text{for} \quad n \geq 1.$$

Since $c_1 = 0$ (the first state visited is $X_0 = 0$), we have $c_n = n(n-1)/2$.

11.2. The Matthews Method

It is not surprising that there is an essentially monotone relationship between hitting times and cover times: the longer it takes to travel between states, the longer it should take to visit all of them. In one direction, it is easy to write down a bound. Fix an irreducible chain with state space Ω. Recall the definition (10.5) of t_{hit}, and let $x, y \in \Omega$ be states for which $t_{\text{hit}} = \mathbf{E}_x \tau_y$. Since any walk started at

x must have visited y by the time all states are covered, we have

$$t_{\text{hit}} = \mathbf{E}_x \tau_y \leq \mathbf{E}_x \tau_{\text{cov}} \leq t_{\text{cov}}. \tag{11.2}$$

It is more interesting to give an upper bound on cover times in terms of hitting times. A walk covering all the states can visit them in many different orders, and this indeterminacy can be exploited. Randomizing the order in which we check whether states have been visited (which, following Aldous and Fill (1999), we will call the Matthews method—see Matthews (1988a) for the original version) allows us to prove both upper and lower bounds. Despite the simplicity of the arguments, these bounds are often remarkably good.

THEOREM 11.2 (Matthews (1988a)). *Let (X_t) be an irreducible finite Markov chain on n states. Then*

$$t_{\text{cov}} \leq t_{\text{hit}} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).$$

PROOF. Without loss of generality, we may assume that our state space is $\{1, \ldots, n\}$. Choose an arbitrary initial state $x \in \Omega$ and let $\sigma \in S_n$ be a uniform random permutation, chosen independently of the chain. We will look for states in order σ. Let T_k be the first time that the states $\sigma(1), \ldots, \sigma(k)$ have all been visited, and let $L_k = X_{T_k}$ be the last state among $\sigma(1), \ldots, \sigma(k)$ to be visited.

Of course, when $\sigma(1) = x$, we have $T_1 = 0$. We will not usually be so lucky. For any $s \in \Omega$, we have

$$\mathbf{E}_x(T_1 \mid \sigma(1) = s) = \mathbf{E}_x(\tau_s) \leq t_{\text{hit}}.$$

Since the events $\{\sigma(1) = s\}$ are disjoint for distinct $s \in \Omega$, Exercise 11.1 ensures that $\mathbf{E}_x(T_1) \leq t_{\text{hit}}$.

How much further along is T_2 than T_1?

- When the chain visits $\sigma(1)$ before $\sigma(2)$, then $T_2 - T_1$ is the time required to travel from $\sigma(1)$ to $\sigma(2)$, and $L_2 = \sigma(2)$.
- When the chain visits $\sigma(2)$ before $\sigma(1)$, we have $T_2 - T_1 = 0$ and $L_2 = \sigma(1)$.

Let's analyze the first case a little more closely. For any two distinct states $r, s \in \Omega$, define the event

$$A_2(r, s) = \{\sigma(1) = r \text{ and } \sigma(2) = L_2 = s\}.$$

Clearly

$$\mathbf{E}_x(T_2 - T_1 \mid A_2(r, s)) = \mathbf{E}_r(\tau_s) \leq t_{\text{hit}}.$$

Conveniently,

$$A_2 = \bigcup_{r \neq s} A_2(r, s)$$

is simply the event that $\sigma(2)$ is visited after $\sigma(1)$, that is, $L_2 = \sigma(2)$. By Exercise 11.1,

$$\mathbf{E}_x(T_2 - T_1 \mid A_2) \leq t_{\text{hit}}.$$

Just as conveniently, A_2^c is the event that $\sigma(2)$ is visited before $\sigma(1)$. It immediately follows that

$$\mathbf{E}_x(T_2 - T_1 \mid A_2^c) = 0.$$

Since σ was chosen uniformly and independently of the chain trajectory, it is equally likely for the chain to visit $\sigma(2)$ before $\sigma(1)$ or after $\sigma(1)$. Thus

$$\begin{aligned}\mathbf{E}_x(T_2 - T_1) &= \mathbf{P}_x(A_2)\mathbf{E}_x(T_2 - T_1 \mid A_2) + \mathbf{P}_x(A_2^c)\mathbf{E}_x(T_2 - T_1 \mid A_2^c) \\ &\leq \frac{1}{2} t_{\text{hit}}.\end{aligned}$$

We estimate $T_k - T_{k-1}$ for $3 \leq k \leq n$ in the same fashion. Now we carefully track whether $L_k = \sigma(k)$ or not. For any distinct $r, s \in \Omega$, define

$$A_k(r,s) = \{\sigma(k-1) = r \text{ and } \sigma(k) = L_k = s\}.$$

Suppose $L_{k-1} = X_{T_k}$ has distribution μ. Then by Exercise 11.1 we have

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k(r,s)) = \mathbf{E}_\mu(\tau_s) = \sum_{i=1}^{n} \mu(i)\mathbf{E}_i(\tau_s) \leq t_{\text{hit}} \tag{11.3}$$

and

$$A_k = \bigcup_{r \neq s} A_k(r,s)$$

is the event that $L_k = \sigma(k)$. Just as above, Exercise 11.1 implies that

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k) \leq t_{\text{hit}},$$

while

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k^c) = 0.$$

Since the permutation σ was chosen both uniformly and independently of the trajectory of the chain, each of $\sigma(1), \ldots, \sigma(k)$ is equally likely to be the last visited. Thus $\mathbf{P}_x(A_k) = 1/k$ and

$$\begin{aligned}\mathbf{E}_x(T_k - T_{k-1}) &= \mathbf{P}_x(A_k)\mathbf{E}_x(T_k - T_{k-1} \mid A_k) + \mathbf{P}_x(A_k^c)\mathbf{E}_x(T_k - T_{k-1} \mid A_k^c) \\ &\leq \frac{1}{k} t_{\text{hit}}.\end{aligned}$$

Finally, summing all these estimates yields

$$\begin{aligned}\mathbf{E}_x(\tau_{\text{cov}}) &= \mathbf{E}_x(T_n) \\ &= \mathbf{E}_x(T_1) + \mathbf{E}_x(T_2 - T_1) + \cdots + \mathbf{E}_x(T_n - T_{n-1}) \\ &\leq t_{\text{hit}} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).\end{aligned}$$

■

EXAMPLE 11.3. The proof above strongly parallels the standard argument for the coupon collecting problem, which we discussed in Section 2.2 and have applied several times: for instance, coupon collector bounds were used to lower bound mixing times for both random walk on the hypercube (Proposition 7.13) and Glauber dynamics on the graph with no edges (Exercise 7.3). For random walk on a complete graph with self-loops, the cover time coincides with the time to "collect" all coupons. In this case $\mathbf{E}_\alpha(\tau_\beta) = n$ is constant for $\alpha \neq \beta$, so the upper bound is tight.

A slight modification of this technique can be used to prove lower bounds: instead of looking for every state along the way to the cover time, we look for the elements of some $A \subseteq \Omega$. Define τ_{cov}^A to be the first time such that every state of A has been visited by the chain. When the elements of A are far away from each other, in the sense that the hitting time between any two of them is large, the time to visit just the elements of A can give a good lower bound on the overall cover time.

PROPOSITION 11.4. *Let $A \subset X$. Set $t_{\text{min}}^A = \min_{a,b \in A, a \neq b} \mathbf{E}_a(\tau_b)$. Then*

$$t_{\text{cov}} \geq \max_{A \subseteq \Omega} t_{\text{min}}^A \left(1 + \frac{1}{2} + \cdots + \frac{1}{|A|-1}\right).$$

PROOF. Fix an initial state $x \in A$ and let σ be a uniform random permutation of the elements of A, chosen independently of the chain trajectory. Let T_k be the first time at which all of $\sigma(1), \sigma(2), \ldots, \sigma(k)$ have been visited, and let $L_k = X_{T_k}$.

With probability $1/|A|$ we have $\sigma(1) = x$ and $T_1 = 0$. Otherwise, the walk must proceed from x to $\sigma(1)$. Thus

$$\mathbf{E}_x(T_1) \geq \frac{1}{|A|} 0 + \frac{|A|-1}{|A|} t_{\text{min}}^A = \left(1 - \frac{1}{|A|}\right) t_{\text{min}}^A. \tag{11.4}$$

For $2 \leq k \leq |A|$ and $r, s \in A$, define

$$B_k(r,s) = \{\sigma(k-1) = r \text{ and } \sigma(k) = L_k = s\},$$

so that, by an argument similar to that of (11.3), using (an obvious corollary to) Exercise 11.1, we have

$$\mathbf{E}_x(T_k - T_{k-1} \mid B_k(r,s)) \geq t_{\text{min}}^A.$$

Then

$$B_k = \bigcup_{r,s \in A} B_k(r,s)$$

is the event that $L_k = \sigma(k)$. Now

$$\mathbf{E}_x(T_k - T_{k-1} \mid B_k^c) = 0 \quad \text{and} \quad \mathbf{E}_x(T_k - T_{k-1} | B_k) \geq t_{\text{min}}^A.$$

By the uniformity and independence of σ we have $\mathbf{P}(B_k) = 1/k$ and thus

$$\mathbf{E}_x(T_k - T_{k-1}) \geq \frac{1}{k} t_{\text{min}}^A. \tag{11.5}$$

Adding up (11.4) and the bound of (11.5) for $2 \leq k \leq |A|$ gives

$$\mathbf{E}_x(\tau_{\text{cov}}^A) \geq t_{\text{min}}^A \left(1 + \frac{1}{2} + \cdots + \frac{1}{|A|-1}\right)$$

(note that the negative portion of the first term cancels with the last term).

Since $t_{\text{cov}} \geq \mathbf{E}_x(\tau_{\text{cov}}) \geq \mathbf{E}_x(\tau_{\text{cov}}^A)$ for every $x \in A$, we are done. ∎

REMARK 11.5. While any subset A yields a lower bound, some choices for A are uninformative. For example, when the underlying graph of (Y_t) contains a leaf, $t_{\text{min}}^A = 1$ for any set A containing both the leaf and its (unique) neighbor.

11.3. Applications of the Matthews Method

11.3.1. Binary trees. Consider simple random walk on the rooted binary tree with depth k and $n = 2^{k+1}-1$ vertices, which we first discussed in Section 5.3.4. The maximal hitting time will be realized by pairs of leaves a, b whose most recent common ancestor is the root (see Exercise 10.4). For such a pair, the hitting time will, by symmetry, be the same as the commute time between the root and one of the leaves. By Proposition 10.6 (the Commute Time Identity), we have

$$\mathbf{E}_a \tau_b = 2(n-1)k$$

(since the effective resistance between the root and the leaf is k, by Example 9.7, and the total conductance c_G of the network is twice the number of edges). Hence Theorem 11.2 gives

$$t_{\text{cov}} \leq 2(n-1)k\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) = (2 + o(1))(\log 2)nk^2. \tag{11.6}$$

What about a lower bound? We need an appropriate set $A \subseteq X$. Fix a level h in the tree, and let A be a set of 2^h leaves chosen so that each vertex at level h has a unique descendant in A. Notice that the larger h is, the more vertices there are in A—and the closer together those vertices can be. We will choose a value of h below to optimize our bound.

Consider two distinct leaves $a, b \in A$. If their closest common ancestor is at level $h' < h$, then the hitting time from one to the other is the same as the commute time from their common ancestor to one of them, say a. Again, by the Commute Time Identity (Proposition 10.6) and Example 9.7, this is exactly

$$\mathbf{E}_a \tau_b = 2(n-1)(k-h'),$$

which is clearly minimized when $h' = h - 1$. By Proposition 11.4,

$$t_{\text{cov}} \geq 2(n-1)(k-h+1)\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^h - 1}\right) = (2 + o(1))(\log 2)n(k-h)h. \tag{11.7}$$

Setting $h = \lfloor k/2 \rfloor$ in (11.7) gives

$$t_{\text{cov}} \geq \frac{1}{4} \cdot (2 + o(1))(\log 2)nk^2. \tag{11.8}$$

There is still a factor of 4 gap between the upper bound of (11.6) and the lower bound of (11.8). In fact, the upper bound is sharp. See the Notes.

11.3.2. Tori. In Section 10.4 we derived fairly sharp (up to constants) estimates for the hitting times of simple random walks on finite tori of various dimensions. Let's use these bounds and the Matthews method to determine equally sharp bounds on the expected cover times of tori. We discuss the case of dimension at least 3 first, since the details are a bit simpler.

When the dimension $d \geq 3$, Proposition 10.13 tells us that there exist constants c_d and C_d such that for any distinct vertices x, y of $\mathbb{Z}_n^d$,

$$c_d n^d \leq \mathbf{E}_x(\tau_y) \leq C_d n^d.$$

Remarkably, this bound does not depend on the distance between x and y! By Theorem 11.2, the average cover time satisfies

$$t_{\text{cov}} \le C_d n^d \left(1 + \frac{1}{2} + \cdots + \frac{1}{n^d}\right) \tag{11.9}$$

$$= C_d d n^d \log n(1 + o(1)). \tag{11.10}$$

To derive an almost-matching lower bound from Proposition 11.4, we must choose a set A large enough that the sum of reciprocals in the second factor is substantial. Let's take A to be $\mathbb{Z}_n^d$ itself (any set containing a fixed positive fraction of the points of the torus would work as well). Then

$$t_{\text{cov}} \ge t_{\min}^A \left(1 + \frac{1}{2} + \cdots + \frac{1}{|A| - 1}\right)$$

$$\ge c_d d n^d \log n(1 + o(1)),$$

which is only a constant factor away from our upper bound.

In dimension 2, Proposition 10.13 says that when x and y are vertices of $\mathbb{Z}_n^2$ at distance k,

$$c_2 n^2 \log(k) \le \mathbf{E}_x(\tau_y) \le C_2 n^2 \log(k).$$

In this case the Matthews upper bound gives

$$\mathbf{E}(\tau_{\text{cov}}) \le 2C_2 n^2 (\log n)^2 (1 + o(1)), \tag{11.11}$$

since the furthest apart two points can be is n.

To get a good lower bound, we must choose a set A which is as large as possible, but for which the minimum distance between points is also large. Assume for simplicity that n is a perfect square, and let A be the set of all points in Z_n^2 both of whose coordinates are multiples of $\sqrt{n}$. Then Proposition 11.4 and Proposition 10.13 imply

$$\mathbf{E}(\tau_{\text{cov}}) \ge c_2 n^2 \log(\sqrt{n}) \left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1}\right)$$

$$= \frac{c_2}{2} n^2 (\log n)^2 (1 + o(1)).$$

Figure 11.1 shows the points of a 75×75 torus left uncovered by a single random walk trajectory at equally spaced fractions of its cover time.

Exercises 11.4 and 11.5 use improved estimates on the hitting times to get our upper and lower bounds for cover times on tori even closer together.

11.3.3. Waiting for all patterns in coin tossing. In Section 17.3.2, we will use elementary martingale methods to compute the expected time to the first occurrence of a specified pattern (such as $HTHHTTH$) in a sequence of independent fair coin tosses. Here we examine the time required for *all* 2^k patterns of length k to have appeared. In order to apply the Matthews method, we first give a simple universal bound on the expected hitting time of any pattern.

Consider the Markov chain whose state space is the collection $\Omega = \{0,1\}^k$ of binary k-tuples and whose transitions are as follows: at each step, delete the leftmost bit and append on the right a new fair random bit independent of all earlier bits. We can also view this chain as sliding a window of width k from left to right along a stream of independent fair bits. (In fact, the winning streak chain of

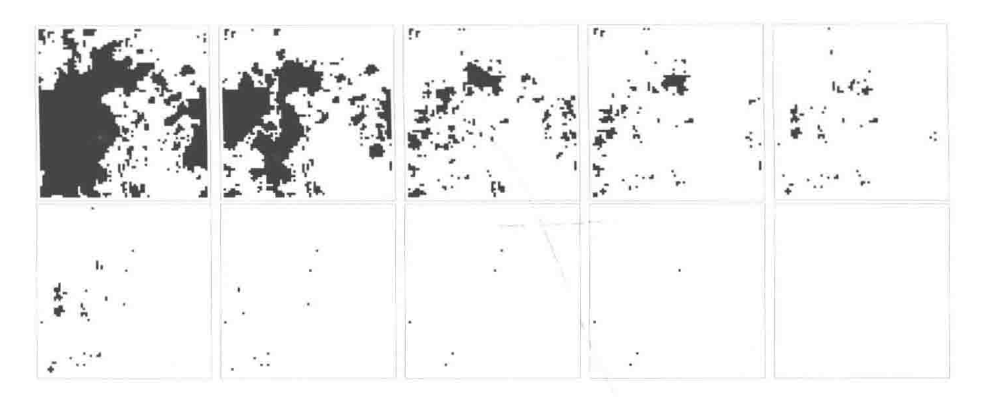


FIGURE 11.1. Black squares show the states unvisited by a single trajectory of simple random walk on a 75×75 torus. This trajectory took 145,404 steps to cover. The diagrams show the walk after 10%, 20%, ..., 100% of its cover time.

Example 4.15 is a lumping of this chain—see Lemma 2.5.) We call this the ***shift chain on binary k-tuples***.

In the coin tossing picture, it is natural to consider the *waiting time* w_x for a pattern $x \in \{0,1\}^k$, which is defined to be the number of steps required for x to appear using all "new" bits—that is, without any overlap with the initial state. Note that

$$w_x \geq k \quad \text{and} \quad w_x \geq \tau_x \quad \text{for all } x \in \{0,1\}^k. \tag{11.12}$$

Also, w_x does not depend on the initial state of the chain. Hence

$$\mathbf{E}w_x \geq \mathbf{E}_x\tau_x^+ = 2^k \tag{11.13}$$

(the last equality follows immediately from (1.26), since our chain has a uniform stationary distribution).

LEMMA 11.6. *Fix $k \geq 1$. For the shift chain on binary k-tuples,*

$$H_k := \max_{x \in \{0,1\}^k} \mathbf{E}w_x = 2^{k+1} - 2.$$

PROOF. When $k = 1$, w_x is geometric with parameter 2. Hence $H_1 = 2$.

Now fix a pattern x of length $k+1$ and let x^- be the pattern consisting of the first k bits of x. To arrive at x, we must first build up x^-. Flipping one more coin has probability $1/2$ of completing pattern x. If it does not, we resume waiting for x. The additional time required is certainly bounded by the time required to construct x from entirely new bits. Hence

$$\mathbf{E}w_x \leq \mathbf{E}w_{x^-} + 1 + \frac{1}{2}\mathbf{E}w_x. \tag{11.14}$$

To bound H_{k+1} in terms of H_k, choose an x that achieves $H_{k+1} = \mathbf{E}w_x$. On the right-hand-side of (11.14), the first term is bounded by H_k, while the third is equal to $(1/2)H_{k+1}$. We conclude that

$$H_{k+1} \leq H_k + 1 + \frac{1}{2}H_{k+1},$$

which can be rewritten as

$$H_{k+1} \leq 2H_k + 2.$$

This recursion, together with the initial condition $H_1 = 2$, implies $H_k \le 2^{k+1} - 2$.

When x is a constant pattern (all 0's or all 1's) of length k and y is any pattern ending in the opposite bit, we have $\mathbf{E}_y \tau_x = H_k = 2^{k+1} - 2$. Indeed, since one inappropriate bit requires a copy of x to be built from new bits, equality of hitting time and waiting time holds throughout the induction above. ∎

We can now combine Lemma 11.6 with (11.12) and the Matthews upper bound of Theorem 11.2, obtaining

$$\mathbf{E}_x(\tau_{\text{cov}}) \le H_k \left(1 + \frac{1}{2} + \cdots + \frac{1}{2^k}\right) = (\log 2) k 2^{k+1}(1 + o(1)).$$

Looking more closely at the relationship between hitting times and waiting times will allow us to improve this upper bound by a factor of 2 and to prove a matching lower bound.

LEMMA 11.7. *Let $\theta = \theta_{a,b} = \mathbf{P}_a(\tau_b^+ < k)$. Then for any $a, b \in \{0,1\}^k$ we have*

$$\mathbf{E} w_b \le \frac{\mathbf{E}_a \tau_b^+ + k\theta}{1 - \theta}.$$

PROOF. The following inequality is true:

$$w_b \le \tau_b^+ + \mathbf{1}_{\{\tau_b^+ < k\}}(k + w_b^*), \tag{11.15}$$

where w_b^* is the amount of time required to build b with all new bits, starting after the k-th bit has been added. (Note that w_b^* has the same distribution as w_b.) Why? When $\tau_b^+ \ge k$, we have $w_b = \tau_b^+$. When $\tau_b^+ < k$, we can ensure a copy of b made from new bits by waiting for k bits, then restarting our counter and waiting for a copy of b consisting of bits added after the first k.

Since w_b^* in independent of the event $\{\tau_b^+ < k\}$, taking expectations on both sides of (11.15) yields

$$\mathbf{E} w_b \le \mathbf{E}_a \tau_b^+ + \theta(k + \mathbf{E} w_b)$$

(since $\mathbf{E}_a w_b$ does not depend on the initial state a, we drop the subscript), and rearranging terms completes the proof. ∎

PROPOSITION 11.8. *The cover time satisfies*

$$t_{\text{cov}} \ge (\log 2) k 2^k (1 + o(1)).$$

PROOF. Fix $j = \lceil \log_2 k \rceil$ and let $A \subseteq \{0,1\}^k$ consist of those bitstrings that end with j zeroes followed by a 1. Fix $a, b \in A$, where $a \ne b$. By Lemma 11.7, we have

$$\mathbf{E}_a \tau_b^+ > (1 - \theta)\mathbf{E} w_b - k\theta,$$

where our choice of A ensures

$$\theta = \mathbf{P}_a(\tau_b^+ < k) < 2^{-(j+1)} + \cdots + 2^{-(k-1)} < 2^{-j}.$$

By (11.13) we may conclude

$$\mathbf{E}_a \tau_b^+ > 2^k (1 + o(1)).$$

Now apply Proposition 11.4. Since $|A| = 2^{k-j-1}$, we get

$$t_{\text{cov}} \ge (k - j - 1)(\log 2) 2^k (1 + o(1)) = (\log 2) k 2^k (1 + o(1)).$$

∎

To improve the upper bound, we apply a variant on the Matthews method which, at first glance, may seem unlikely to help. For any $B \subseteq \Omega$, the argument for the Matthews bound immediately gives

$$\mathbf{E}_x \tau_{\rm cov}^B \le \max_{b,b' \in B} \mathbf{E}_b \tau_b' \left(1 + \frac{1}{2} + \cdots + \frac{1}{|B|}\right). \tag{11.16}$$

Certainly the total cover time $\tau_{\rm cov}$ is bounded by the time taken to visit first all the states in B and then all the states in B^c. Hence

$$\mathbf{E}_x \tau_{\rm cov} \le \mathbf{E}_x \tau_{\rm cov}^B + \max_{y \in \Omega} \mathbf{E}_y \tau_{\rm cov}^{B^c}. \tag{11.17}$$

If the states that take a long time to hit form a small fraction of Ω, then separating those states from the rest can yield a better bound on $t_{\rm cov}$ than direct application of Theorem 11.2. For the current example of waiting for all possible patterns in coin tossing, we improve the bound by a factor of 2—obtaining an asymptotic match with the lower bound of Proposition 11.8.

PROPOSITION 11.9. *The cover time satisfies*

$$t_{\rm cov} \le (\log 2) k 2^k (1 + o(1)).$$

PROOF. We partition the state space $\{0,1\}^k$ into two sets. Fix $j = \lceil \log_2 k \rceil$ and let B be the set of all strings $b \in \{0,1\}^k$ with the following property: any bitstring that is both a suffix and a prefix of b must have length less than $k - j$. (In other words, elements of B must be shifted by *more than* j bits in order to agree with themselves. For any string $b \in B$, we must have $\tau_b^+ > j$.)

Since for $m < k$ there are only 2^m strings of length k that agree with themselves after shifting by m bits, we have $|B^c| = 2 + 4 + \cdots + 2^j \le 2^{j+1} \le 4k$.

For $a, b \in B$, we can use Lemma 11.7 to bound the maximum expected hitting time. We have

$$\mathbf{E}_a \tau_b \le \mathbf{E} w_b \le \frac{\mathbf{E}_b \tau_b^+ + k\theta}{1 - \theta}.$$

(Since $\mathbf{E} w_b$ does not depend on the initial state, we have taken the initial state to be b as we apply Lemma 11.7.)

Since our chain has a uniform stationary distribution, (1.26) implies that $\mathbf{E}_b \tau_b^+ = 2^k$. By our choice of B, we have $\theta = \mathbf{P}_b(\tau_b^+ < k) < 1/k$. Thus

$$\mathbf{E}_a \tau_b \le \frac{2^k + k(1/k)}{1 - 1/k} = 2^k (1 + o(1)). \tag{11.18}$$

For $a, b \in B^c$, we again use Lemma 11.6 to bound $\mathbf{E}_a \tau_b$. Finally we apply (11.17), obtaining

$$\begin{aligned} t_{\rm cov} &\le (\log |B| + o(1)) \left(2^k (1 + o(1))\right) + (\log |B^c| + o(1)) \left(2^{k+1} + o(1)\right)) \\ &= (\log 2) k 2^k (1 + o(1)). \end{aligned}$$

■

Exercises

EXERCISE 11.1. Let Y be a random variable on some probability space, and let $B = \bigcup_j B_j$ be a partition of an event B into (finitely or countably many) disjoint subevents B_j.

(a) Prove that when $\mathbf{E}(Y \mid B_j) \le M$ for every j, then $\mathbf{E}(Y \mid B) \le M$.

(b) Give an example to show that the conclusion of part (a) can fail when the events B_j are not disjoint.

EXERCISE 11.2. What upper and lower bounds does the Matthews method give for cycle $\mathbb{Z}_n$? Compare to the actual value, computed in Example 11.1, and explain why the Matthews method gives a poor result for this family of chains.

EXERCISE 11.3. Show that the cover time of the m-dimensional hypercube is asymptotic to $m2^m \log(2)$ as $m \to \infty$.

EXERCISE 11.4. In this exercise, we demonstrate that for tori of dimension $d \geq 3$, just a little more information on the hitting times suffices to prove a matching lower bound.

(a) Show that when a sequence of pairs of points $x_n, y_n \in \mathbb{Z}_n^d$ has the property that the distance between them tends to infinity with n, then the upper-bound constant C_d of (10.18) can be chosen so that $\mathbf{E}_{x_n}(\tau_{y_n})/n^d \to C_d$.
(b) Give a lower bound on t_{cov} that has the same initial constant as the upper bound of (11.9).

EXERCISE 11.5. Following the example of Exercise 11.4, derive a lower bound for $\mathbf{E}(\tau_{\text{cov}})$ on the two-dimensional torus that is within a factor of 4 of the upper bound (11.11).

Notes

The Matthews method first appeared in Matthews (1988a). Matthews (1989) looked at the cover time of the hypercube, which appears in Exercise 11.3.

The argument we give for a lower bound on the cover time of the binary tree is due to Zuckerman (1992). Aldous (1991a) shows that the upper bound is asymptotically sharp; Peres (2002) presents a simpler version of the argument.

In the *American Mathematical Monthly*, Herb Wilf (1989) described his surprise at the time required for a simulated random walker to visit every pixel of his computer screen. This time is, of course, the cover time for the two-dimensional finite torus. The exact asymptotics of the expected cover time on Z_n^2 have been determined. Zuckerman (1992) estimated the expected cover time to within a constant, while Dembo, Peres, Rosen, and Zeitouni (2004) showed that

$$\mathbf{E}(\tau_{\text{cov}}) \sim \frac{4}{\pi} n^2 (\log n)^2.$$

Móri (1987) found the cover time for all patterns of length k using ideas from Aldous (1983a). The collection Godbole and Papastavridis (1994) has many further papers on this topic. A single issue of the *Journal of Theoretical Probability* contained several papers on cover times: these include Aldous (1989a), Aldous (1989b), Broder and Karlin (1989), Kahn, Linial, Nisan, and Saks (1989), and Zuckerman (1989).

Aldous (1991b) gives a condition guaranteeing that the cover time is well-approximated by its expected value. See Theorem 19.8 for a statement.

CHAPTER 12

Eigenvalues

12.1. The Spectral Representation of a Reversible Transition Matrix

We begin by collecting some elementary facts about the eigenvalues of transition matrices, which we leave to the reader to verify (Exercise 12.1):

LEMMA 12.1. *Let P be the transition matrix of a finite Markov chain.*

(i) *If λ is an eigenvalue of P, then $|\lambda| \leq 1$.*
(ii) *If P is irreducible, the vector space of eigenfunctions corresponding to the eigenvalue 1 is the one-dimensional space generated by the column vector $\mathbf{1} := (1, 1, \ldots, 1)^T$.*
(iii) *If P is irreducible and aperiodic, then -1 is not an eigenvalue of P.*

Denote by $\langle \cdot, \cdot \rangle$ the usual inner product on $\mathbb{R}^\Omega$, given by $\langle f, g \rangle = \sum_{x \in \Omega} f(x)g(x)$. We will also need another inner product, denoted by $\langle \cdot, \cdot \rangle_\pi$ and defined by

$$\langle f, g \rangle_\pi := \sum_{x \in \Omega} f(x)g(x)\pi(x). \tag{12.1}$$

We write $\ell^2(\pi)$ for the vector space $\mathbb{R}^\Omega$ equipped with the inner product (12.1). Because we regard elements of $\mathbb{R}^\Omega$ as functions from Ω to $\mathbb{R}$, we will call eigenvectors of the matrix P eigenfunctions.

Recall that the transition matrix P is reversible with respect to the stationary distribution π if $\pi(x)P(x, y) = \pi(y)P(y, x)$ for all $x, y \in \Omega$. The reason for introducing the inner product (12.1) is

LEMMA 12.2. *Let P be reversible with respect to π.*

(i) *The inner product space $(\mathbb{R}^\Omega, \langle \cdot, \cdot \rangle_\pi)$ has an orthonormal basis of real-valued eigenfunctions $\{f_j\}_{j=1}^{|\Omega|}$ corresponding to real eigenvalues $\{\lambda_j\}$.*
(ii) *The matrix P can be decomposed as*

$$\frac{P^t(x, y)}{\pi(y)} = \sum_{j=1}^{|\Omega|} f_j(x) f_j(y) \lambda_j^t.$$

(iii) *The eigenfunction f_1 corresponding to the eigenvalue 1 can be taken to be the constant vector $\mathbf{1}$, in which case*

$$\frac{P^t(x, y)}{\pi(y)} = 1 + \sum_{j=2}^{|\Omega|} f_j(x) f_j(y) \lambda_j^t. \tag{12.2}$$

PROOF. Define $A(x, y) := \pi(x)^{1/2}\pi(y)^{-1/2}P(x, y)$. Reversibility of P implies that A is symmetric. The spectral theorem for symmetric matrices (Theorem A.11) guarantees that the inner product space $(\mathbb{R}^\Omega, \langle \cdot, \cdot \rangle)$ has an orthonormal basis $\{\varphi_j\}_{j=1}^{|\Omega|}$ such that φ_j is an eigenfunction with real eigenvalue λ_j.

The reader should directly check that $\sqrt{\pi}$ is an eigenfunction of A with corresponding eigenvalue 1; we set $\varphi_1 := \sqrt{\pi}$ and $\lambda_1 := 1$.

If D_π denotes the diagonal matrix with diagonal entries $D_\pi(x,x) = \pi(x)$, then $A = D_\pi^{\frac{1}{2}} P D_\pi^{-\frac{1}{2}}$. If $f_j := D_\pi^{-\frac{1}{2}}\varphi_j$, then f_j is an eigenfunction of P with eigenvalue λ_j:

$$Pf_j = PD_\pi^{-\frac{1}{2}}\varphi_j = D_\pi^{-\frac{1}{2}}(D_\pi^{\frac{1}{2}}PD_\pi^{-\frac{1}{2}})\varphi_j = D_\pi^{-\frac{1}{2}}A\varphi_j = D_\pi^{-\frac{1}{2}}\lambda_j\varphi_j = \lambda_j f_j.$$

Although the eigenfunctions $\{f_j\}$ are not necessarily orthonormal with respect to the usual inner product, they are orthonormal with respect to the inner product $\langle\cdot,\cdot\rangle_\pi$ defined in (12.1):

$$\delta_{ij} = \langle\varphi_i,\varphi_j\rangle = \langle D_\pi^{\frac{1}{2}}f_i, D_\pi^{\frac{1}{2}}f_j\rangle = \langle f_i, f_j\rangle_\pi. \tag{12.3}$$

(The first equality follows since $\{\varphi_j\}$ is orthonormal with respect to the usual inner product.) This proves (i).

Let δ_y be the function

$$\delta_y(x) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{if } y \neq x. \end{cases}$$

Considering $(\mathbb{R}^\Omega, \langle\cdot,\cdot\rangle_\pi)$ with its orthonormal basis of eigenfunctions $\{f_j\}_{j=1}^{|\Omega|}$, the function δ_y can be written via basis decomposition as

$$\delta_y = \sum_{j=1}^{|\Omega|}\langle\delta_y, f_j\rangle_\pi f_j = \sum_{j=1}^{|\Omega|} f_j(y)\pi(y)f_j. \tag{12.4}$$

Since $P^t f_j = \lambda_j^t f_j$ and $P^t(x,y) = (P^t\delta_y)(x)$,

$$P^t(x,y) = \sum_{j=1}^{|\Omega|} f_j(y)\pi(y)\lambda_j^t f_j(x).$$

Dividing by $\pi(y)$ completes the proof of (ii), and (iii) follows from observations above. ■

It follows from Lemma 12.2 that for a function $f : \Omega \to \mathbb{R}$,

$$P^t f = \sum_{j=1}^{|\Omega|}\langle f, f_j\rangle_\pi f_j \lambda_j^t. \tag{12.5}$$

12.2. The Relaxation Time

For a reversible transition matrix P, we label the eigenvalues of P in decreasing order:

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} \geq -1. \tag{12.6}$$

Define

$$\lambda_\star := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P,\ \lambda \neq 1\}. \tag{12.7}$$

The difference $\gamma_\star := 1 - \lambda_\star$ is called the ***absolute spectral gap***. Lemma 12.1 implies that if P is aperiodic and irreducible, then $\gamma_\star > 0$.

The ***spectral gap*** of a reversible chain is defined by $\gamma := 1 - \lambda_2$. Exercise 12.3 shows that if the chain is lazy, then $\gamma_\star = \gamma$.

The ***relaxation time*** t_{rel} of a reversible Markov chain with absolute spectral gap $\gamma_\star$ is defined to be

$$t_{\mathrm{rel}} := \frac{1}{\gamma_\star}.$$

One operational meaning of the relaxation time comes from the inequality

$$\mathrm{Var}_\pi(P^t f) \le (1-\gamma_\star)^{2t} \,\mathrm{Var}_\pi(f). \tag{12.8}$$

(Exercise 12.4 asks for a proof.) By the Convergence Theorem (Theorem 4.9), $P^t f(x) \to E_\pi(f)$ for any $x \in \Omega$, i.e., the function $P^t f$ approaches a constant function. Using (12.8), we can make a quantitative statement: if $t \ge t_{\mathrm{rel}}$, then the standard deviation of $P^t f$ is bounded by $1/e$ times the standard deviation of f. Let $i_\star$ be the value for which $|\lambda_{i_\star}|$ is maximized. Then equality in (12.8) is achieved for $f = f_{i_\star}$, whence the inequality is sharp.

We prove both upper and lower bounds on the mixing time in terms of the relaxation time and the stationary distribution of the chain.

THEOREM 12.3. *Let P be the transition matrix of a reversible, irreducible Markov chain with state space Ω, and let $\pi_{\min} := \min_{x\in\Omega} \pi(x)$. Then*

$$t_{\mathrm{mix}}(\varepsilon) \le \log\left(\frac{1}{\varepsilon \pi_{\min}}\right) t_{\mathrm{rel}}. \tag{12.9}$$

PROOF. Using (12.2) and applying the Cauchy-Schwarz inequality yields

$$\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \le \sum_{j=2}^{|\Omega|} |f_j(x) f_j(y)| \lambda_\star^t \le \lambda_\star^t \left[\sum_{j=2}^{|\Omega|} f_j^2(x) \sum_{j=2}^{|\Omega|} f_j^2(y)\right]^{1/2}. \tag{12.10}$$

Using (12.4) and the orthonormality of $\{f_j\}$ shows that

$$\pi(x) = \langle \delta_x, \delta_x \rangle_\pi = \left\langle \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j, \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j \right\rangle_\pi = \pi(x)^2 \sum_{j=1}^{|\Omega|} f_j(x)^2.$$

Consequently, $\sum_{j=2}^{|\Omega|} f_j(x)^2 \le \pi(x)^{-1}$. This bound and (12.10) imply that

$$\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \le \frac{\lambda_\star^t}{\sqrt{\pi(x)\pi(y)}} \le \frac{\lambda_\star^t}{\pi_{\min}} = \frac{(1-\gamma_\star)^t}{\pi_{\min}} \le \frac{e^{-\gamma_\star t}}{\pi_{\min}}. \tag{12.11}$$

Applying Lemma 6.13 shows that $d(t) \le \pi_{\min}^{-1} \exp(-\gamma_\star t)$. The conclusion now follows from the definition of $t_{\mathrm{mix}}(\varepsilon)$. ■

THEOREM 12.4. *For a reversible, irreducible, and aperiodic Markov chain,*

$$t_{\mathrm{mix}}(\varepsilon) \ge (t_{\mathrm{rel}} - 1) \log\left(\frac{1}{2\varepsilon}\right). \tag{12.12}$$

REMARK 12.5. If the absolute spectral gap $\gamma_\star$ is small because the smallest eigenvalue $\lambda_{|\Omega|}$ is near -1, but the spectral gap γ is not small, the slow mixing suggested by this lower bound can be rectified by passing to a lazy chain to make the eigenvalues positive.

PROOF. Suppose that f is an eigenfunction of P with eigenvalue $\lambda \neq 1$, so that $Pf = \lambda f$. Since the eigenfunctions are orthogonal with respect to $\langle \cdot, \cdot \rangle_\pi$ and $\mathbf{1}$ is an eigenfunction, $\sum_{y \in \Omega} \pi(y) f(y) = \langle \mathbf{1}, f \rangle_\pi = 0$. It follows that

$$|\lambda^t f(x)| = |P^t f(x)| = \left| \sum_{y \in \Omega} \left[P^t(x,y) f(y) - \pi(y) f(y) \right] \right| \leq \|f\|_\infty 2d(t).$$

With this inequality, we can obtain a lower bound on the mixing time. Taking x with $|f(x)| = \|f\|_\infty$ yields

$$|\lambda|^t \leq 2d(t). \tag{12.13}$$

Therefore, $|\lambda|^{t_{\text{mix}}(\varepsilon)} \leq 2\varepsilon$, whence

$$t_{\text{mix}}(\varepsilon) \left(\frac{1}{|\lambda|} - 1 \right) \geq t_{\text{mix}}(\varepsilon) \log \left(\frac{1}{|\lambda|} \right) \geq \log \left(\frac{1}{2\varepsilon} \right).$$

Minimizing the left-hand side over eigenvalues different from 1 and rearranging finishes the proof. ■

COROLLARY 12.6. *For a reversible, irreducible, and aperiodic Markov chain,*

$$\lim_{t \to \infty} d(t)^{1/t} = \lambda_\star.$$

PROOF. One direction is immediate from (12.13), and the other follows from (12.11). ■

EXAMPLE 12.7 (Relaxation time of random transpositions). By Corollary 8.10 and Proposition 8.11, we know that for the random transpositions chain on n cards,

$$t_{\text{mix}} = \Theta(n \log n).$$

Hence $t_{\text{rel}} = O(n \log n)$. The stationary distribution is uniform on $\mathcal{S}_n$. Since Stirling's Formula implies $\log(n!) \sim n \log n$, Theorem 12.3 gives only a constant lower bound. In fact, the relaxation time is known (through other methods) to be exactly $n/2$. See Diaconis (1988).

12.3. Eigenvalues and Eigenfunctions of Some Simple Random Walks

Simple random walk on the n-cycle was introduced in Example 1.4. In Example 2.10, we noted that it can be viewed as a random walk on an n-element cyclic group. Here we use that interpretation to find the eigenvalues and eigenfunctions of this chain and some closely related chains.

12.3.1. The cycle. Let $\omega = e^{2\pi i/n}$. In the complex plane, the set $W_n := \{\omega, \omega^2, \ldots, \omega^{n-1}, 1\}$ of the ***n*-th roots of unity** forms a regular n-gon inscribed in the unit circle. Since $\omega^n = 1$, we have

$$\omega^j \omega^k = \omega^{k+j} = \omega^{k+j \bmod n}.$$

Hence $(W_n, \cdot)$ is a cyclic group of order n, generated by ω. In this section, we view simple random walk on the n-cycle as the random walk on the (multiplicative) group W_n with increment distribution uniform on $\{\omega, \omega^{-1}\}$. Let P be the transition matrix of this walk. Every (possibly complex-valued) eigenfunction f of P satisfies

$$\lambda f(\omega^k) = Pf(\omega^k) = \frac{f(\omega^{k-1}) + f(\omega^{k+1})}{2}$$

for $0 \leq k \leq n-1$.

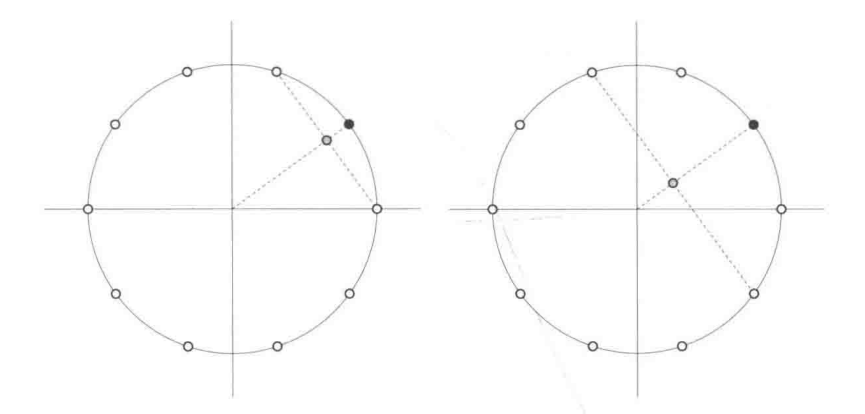


FIGURE 12.1. For simple random walk on the cycle, the eigenvalues must be the cosines. Here $n = 10$. The black vertices represent $\omega = e^{2\pi i/10}$, while the grey vertices represent $(1/2)(\omega^2 + \omega^0)$ and $(1/2)(\omega^3 + \omega^{-1})$, respectively.

For $0 \le j \le n-1$, define $\varphi_j(\omega^k) := \omega^{kj}$. Then

$$P\varphi_j(\omega^k) = \frac{\varphi_j(\omega^{k-1}) + \varphi_j(\omega^{k+1})}{2} = \frac{\omega^{jk+j} + \omega^{jk-j}}{2} = \omega^{jk}\left(\frac{\omega^j + \omega^{-j}}{2}\right). \quad (12.14)$$

Hence φ_j is an eigenfunction of P with eigenvalue $\frac{\omega^j+\omega^{-j}}{2} = \cos(2\pi j/n)$. What is the underlying geometry? As Figure 12.1 illustrates, for any ℓ and j the average of the vectors $\omega^{\ell-j}$ and $\omega^{\ell+j}$ is a scalar multiple of ω^ℓ. Since the chord connecting $\omega^{\ell+j}$ with $\omega^{\ell-j}$ is perpendicular to ω^ℓ, the projection of $\omega^{\ell+j}$ onto ω^ℓ has length $\cos(2\pi j/n)$.

Because φ_j is an eigenfunction of the real matrix P with a real eigenvalue, both its real part and its imaginary parts are eigenfunctions. In particular, the function $f_j : W_n \to \mathbb{R}$ defined by

$$f_j(\omega^k) = \mathrm{Re}(\varphi_j(\omega^k)) = \mathrm{Re}(e^{2\pi ijk/n}) = \cos\left(\frac{2\pi jk}{n}\right) \quad (12.15)$$

is an eigenfunction. We note for future reference that f_j is invariant under complex conjugation of the states of the chain.

We have $\lambda_2 = \cos(2\pi/n) = 1 - \frac{4\pi^2}{2n^2} + O(n^{-4})$, so the spectral gap γ is of order n^{-2} and the relaxation time is of order n^2.

When $n = 2m$ is even, $\cos(2\pi m/n) = -1$ is an eigenvalue, so $\gamma_\star = 0$. The walk in this case is periodic, as we pointed out in Example 1.8.

12.3.2. Lumped chains and the path. Consider the projection of simple random walk on the n-th roots of unity, as described in the preceding section, onto the real axis. The resulting process can take values on a discrete set of points. At most of them (ignoring for the moment those closest to 1 and -1), it is equally likely to move to the right or to the left—just like random walk on the path. Using this idea, we can determine the eigenvalues and eigenfunctions of the random walk on a path with either reflecting boundary conditions or an even chance of holding at the endpoints. First, we give a general lemma on the eigenvalues and eigenfunctions of projected chains (defined in Section 2.3.1).

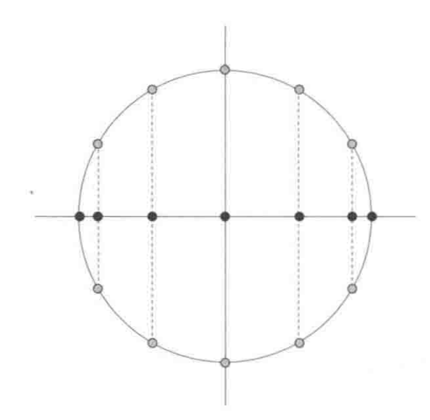


FIGURE 12.2. A random walk on the 12-cycle projects to a random walk on the 7-path. This random walk is reflected when it hits an endpoint.

LEMMA 12.8. *Let Ω be the state space of a Markov chain (X_t) with transition matrix P. Let $\sim$ be an equivalence relation on Ω with equivalence classes $\Omega^\sharp = \{[x] : x \in \Omega\}$ such that $[X_t]$ is a Markov chain with transition matrix $P^\sharp([x],[y]) = P(x,[y])$. Then:*

(i) *Let $f : \Omega \to \mathbb{R}$ be an eigenfunction of P with eigenvalue λ which is constant on each equivalence class. Then the natural projection $f^\sharp : \Omega^\sharp \to R$ of f, defined by $f^\sharp([x]) = f(x)$, is an eigenfunction of $P^\sharp$ with eigenvalue λ.*
(ii) *Conversely, if $g : \Omega^\sharp \to \mathbb{R}$ is an eigenfunction of $P^\sharp$ with eigenvalue λ, then its lift $g^\flat : \Omega \to \mathbb{R}$, defined by $g^\flat(x) = g([x])$, is an eigenfunction of P with eigenvalue λ.*

PROOF. For the first assertion, we can simply compute

$$(Pf^\sharp)([x]) = \sum_{[y]\in\Omega'} P^\sharp([x],[y])f^\sharp([y]) = \sum_{[y]\in\Omega'} P(x,[y])f(y)$$
$$= \sum_{[y]\in\Omega'}\sum_{z\in[y]} P(x,z)f(z) = \sum_{z\in\Omega} P(x,z)f(z) = (Pf)(x) = \lambda f(x) = \lambda f([x]).$$

To prove the second assertion, just run the computations in reverse:

$$(Pf)(x) = \sum_{z\in\Omega} P(x,z)f(z) = \sum_{[y]\in\Omega'}\sum_{z\in[y]} P(x,z)f(z) = \sum_{[y]\in\Omega'} P(x,[y])f(y)$$
$$= \sum_{[y]\in\Omega'} P^\sharp([x],[y])f^\sharp([y]) = (P^\sharp f^\sharp)([x]) = \lambda f^\sharp([x]) = \lambda f(x).$$

■

EXAMPLE 12.9 (Path with reflection at the endpoints). Let $\omega = e^{\pi i/(n-1)}$ and let P be the transition matrix of simple random walk on the $2(n-1)$-cycle identified with random walk on the multiplicative group $W_{2(n-1)} = \{\omega, \omega^2, \ldots, \omega^{2n-1} = 1\}$, as in Section 12.3.1. Now declare $\omega^k \in W_{2(n-1)}$ to be equivalent to its conjugate ω^{-k}. This equivalence relation is compatible with the transitions in the sense required by Lemma 2.5. If we identify each equivalence class with the common

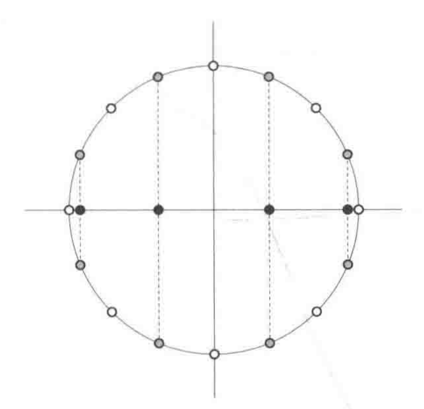


FIGURE 12.3. A random walk on the "odd" states of a 16-cycle projects to a random walk on the 4-path. This lumped walk has holding probability 1/2 at the endpoints of the path.

projection $v_k = \cos(\pi k/(n-1))$ of its elements onto the real axis, the lumped chain is a simple random walk on the path with n vertices $W^\sharp = \{v_0, v_1, \ldots, v_{n-1}\}$ and reflecting boundary conditions. That is, when the walk is at v_0, it moves to v_1 with probability 1 and when the walk is at v_{n-1}, it moves to v_{n-2} with probability 1. (See Figure 12.2.)

By Lemma 2.5 and (12.14), the functions $f_j^\sharp : W^\sharp \to \mathbb{R}$ defined by

$$f_j^\sharp(v_k) = \cos\left(\frac{\pi j k}{(n-1)}\right) \tag{12.16}$$

for $0 \le j \le n-1$ are eigenfunctions of the projected walk. The eigenfunction $f_j^\sharp$ has eigenvalue $\cos(\pi j/(n-1))$. Since we obtain n linearly independent eigenfunctions for n distinct eigenvalues, the functions in (12.16) form a basis.

EXAMPLE 12.10 (Path with holding probability 1/2 at endpoints). Let $\omega = e^{\pi i/(2n)}$. We consider simple random walk on the cycle of length $2n$, realized as a multiplicative random walk on the $2n$-element set

$$W_{\text{odd}} = \{\omega, \omega^3, \ldots, \omega^{4n-1}\}$$

that at each step multiplies the current state by a uniformly chosen element of $\{\omega^2, \omega^{-2}\}$.

Note that this walk is nearly identical to standard simple random walk on the $2n$-th roots of unity; we have rotated the state space through an angle of $\pi/(2n)$, or, equivalently, multiplied each state by ω. The analysis of Section 12.3.1 still applies, so that the function $f_j : W_{\text{odd}} \to \mathbb{R}$ defined by

$$f_j(\omega^{2k+1}) = \cos\left(\frac{\pi(2k+1)j}{2n}\right) \tag{12.17}$$

is an eigenfunction with eigenvalue $\cos(\pi j/n)$.

Now declare each $\omega^{2k+1} \in W_{\text{odd}}$ to be equivalent to its conjugate ω^{-2k-1}. This equivalence relation is compatible with the transitions in the sense required by Lemma 2.5. Again identify each equivalence class with the common projection $u_k = \cos(\pi(2k+1)/(2n))$ of its elements onto the real axis. The lumped chain is

a simple random walk on the path with n vertices $W^\sharp = \{u_0, u_1, \ldots, u_{n-1}\}$ and loops at the endpoints. That is, when the walk is at u_0, it moves to u_1 with probability $1/2$ and stays at u_0 with probability $1/2$, and when the walk is at u_{n-1}, it moves to u_{n-2} with probability $1/2$ and stays at u_{n-1} with probability $1/2$. (See Figure 12.3.)

By Lemma 2.5 and (12.17), the functions $f_j^\sharp : W^\sharp \to \mathbb{R}$ defined by

$$f_j^\sharp(w_k) = \cos\left(\frac{\pi(2k+1)j}{2n}\right) \tag{12.18}$$

for $j = 0, \ldots, n-1$ are eigenfunctions of the random walk on the path $W^\sharp$ with holding at the boundary. The eigenvalue of $f_j^\sharp$ is $\cos(\pi j/n)$. These n linearly independent eigenfunctions form a basis.

12.4. Product Chains

For each $j = 1, 2, \ldots, d$, let P_j be an irreducible transition matrix on the state space Ω_j and let π_j be its stationary distribution. Let w be a probability distribution on $\{1, \ldots, d\}$. Consider the chain on $\tilde{\Omega} := \Omega_1 \times \Omega_2 \cdots \times \Omega_d$ which selects at each step a coordinate i according to the distribution w, and then moves only in the i-th coordinate according to the transition matrix P_i. Let $\boldsymbol{x}$ denote the vector $(x_1, \ldots, x_d)$. The transition matrix $\tilde{P}$ for this chain is

$$\tilde{P}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{d} w_j P_j(x_j, y_j) \prod_{i\,:\, i \neq j} \mathbf{1}\{x_i = y_i\}. \tag{12.19}$$

See Exercise 12.7 for a different product chain.

If $f^{(j)}$ is a function on Ω_j for each $j = 1, 2, \ldots, d$, the ***tensor product*** of $\{f^{(j)}\}_{j=1}^d$ is the function on $\tilde{\Omega}$ defined by

$$(f^{(1)} \otimes f^{(2)} \otimes \cdots \otimes f^{(d)})(x_1, \ldots, x_d) := f^{(1)}(x_1) f^{(2)}(x_2) \cdots f^{(d)}(x_d).$$

If each P_j is irreducible, then so is $\tilde{P}$. If we let $\tilde{\pi} := \pi_1 \otimes \cdots \otimes \pi_d$ (regarding π_j as a function on Ω_j), then it is straightforward to verify that $\tilde{\pi}$ is stationary for $\tilde{P}$.

LEMMA 12.11. *Suppose that for each $j = 1, 2, \ldots, d$, the transition matrix P_j on state space Ω_j has eigenfunction $\varphi^{(j)}$ with eigenvalue $\lambda^{(j)}$. Let w be a probability distribution on $\{1, \ldots, d\}$.*

(i) *The function $\tilde{\varphi} := \varphi^{(1)} \otimes \cdots \otimes \varphi^{(d)}$ is an eigenfunction of the transition matrix $\tilde{P}$ defined in (12.19), with eigenvalue $\sum_{j=1}^d w_j \lambda^{(j)}$.*

(ii) *Suppose for each j, the set $\mathcal{B}_j$ is an orthogonal basis in $\ell^2(\pi_j)$. The collection*

$$\tilde{\mathcal{B}} = \{\varphi^{(1)} \otimes \cdots \otimes \varphi^{(d)} \,:\, \varphi^{(i)} \in \mathcal{B}_i\}$$

is a basis for $\ell^2(\pi_1 \times \cdots \times \pi_d)$.

PROOF. Define $\tilde{P}_j$ on $\tilde{\Omega}$ by

$$\tilde{P}_j(\boldsymbol{x}, \boldsymbol{y}) = P_j(x_j, y_j) \prod_{i\,:\, i \neq j} \mathbf{1}\{x_i = y_i\}.$$

This corresponds to the chain on $\tilde{\Omega}$ which always moves in the j-th coordinate according to P_j. It is simple to check that $\tilde{P}_j \tilde{\varphi}(\boldsymbol{x}) = \lambda_j \tilde{\varphi}(\boldsymbol{x})$. From this and noting

that $\tilde{P} = \sum_{j=1}^{d} w_j \tilde{P}_j$, it follows that

$$\tilde{P}\tilde{\varphi}(\boldsymbol{x}) = \sum_{j=1}^{d} w_j \tilde{P}_j \tilde{\varphi}(\boldsymbol{x}) = \left[\sum_{j=1}^{d} w_j \lambda^{(j)}\right] \tilde{\varphi}(\boldsymbol{x}).$$

We now prove part (ii). Let $\tilde{\varphi} := \varphi^{(1)} \otimes \cdots \otimes \varphi^{(d)}$ and $\tilde{\psi} := \psi^{(1)} \otimes \cdots \otimes \psi^{(d)}$, where $\varphi^{(j)}, \psi^{(j)} \in \mathcal{B}_j$ for all j and $\tilde{\varphi} \neq \tilde{\psi}$. Let j_0 be such that $\varphi^{(j_0)} \neq \psi^{(j_0)}$. We have that

$$\langle \tilde{\varphi}, \tilde{\psi} \rangle_{\tilde{\pi}} = \prod_{j=1}^{d} \langle \varphi^{(j)}, \psi^{(j)} \rangle_{\pi_j} = 0,$$

since the j_0-indexed term vanishes. Therefore, the elements of $\tilde{\mathcal{B}}$ are orthogonal. Since there are $|\Omega_1| \times \cdots \times |\Omega_d|$ elements of $\tilde{\mathcal{B}}$, which equals the dimension of $\tilde{X}$, the collection $\tilde{\mathcal{B}}$ is an orthogonal basis for $\ell^2(\tilde{\pi})$. ■

COROLLARY 12.12. *Let γ_j be the spectral gap for P_j. The spectral gap $\tilde{\gamma}$ for the product chain satisfies*

$$\tilde{\gamma} = \min_{1 \leq j \leq d} w_j \gamma_j.$$

PROOF. By Lemma 12.11, the set of eigenvalues is

$$\left\{ \sum_{i=1}^{d} w_i \lambda^{(i)} : \sum_{i=1}^{d} w_i = 1,\ w_i \geq 0,\ \lambda^{(i)} \text{ an eigenvalue of } P_j \right\}. \tag{12.20}$$

Let i_0 be such that $w_{i_0} \lambda^{(i_0)} = \max_{1 \leq i \leq d} w_i \lambda^{(i)}$. The second largest eigenvalue corresponds to taking $\lambda^{(i)} = 1$ for $i \neq i_0$ and $\lambda^{(i_0)} = 1 - \gamma_{i_0}$. ■

We can apply Corollary 12.12 to bound the spectral gap for Glauber dynamics (defined in Section 3.3.2) when π is a product measure:

LEMMA 12.13. *Suppose that $\{V_i\}$ is a partition of a finite set V, the set S is finite, and that π is a probability distribution on S^V satisfying $\pi = \prod_{i=1}^{d} \pi_i$, where π is a probability on S^{V_i}. Let γ be the spectral gap for the Glauber dynamics on S^V for π, and let γ_i be the spectral gap for the Glauber dynamics on S^{V_i} for π_i. If $n = |V|$ and $n_j = |V_j|$, then*

$$\frac{1}{n\gamma} = \max_{1 \leq j \leq d} \frac{1}{n_j \gamma_j}. \tag{12.21}$$

REMARK 12.14. Suppose the graph G can be decomposed into connected components $G_1, \ldots, G_r$ and that π is the Ising model on G. Then $\pi = \prod_{i=1}^{r} \pi_i$, where π_i is the Ising model on G_i. The corresponding statement is also true for the hardcore model and the uniform distribution on proper colorings.

PROOF OF LEMMA 12.13. If $\Omega(x, v) = \{y \in \Omega : y(w) = x(w) \text{ for all } w \neq v\}$, then the transition matrix is given by

$$P(x, y) = \sum_{v \in V} \frac{1}{n} \frac{\pi(y)}{\pi(\Omega(x, v))} \mathbf{1}\{y \in \Omega(x, v)\}.$$

For $x \in S^V$, write x_i for the projection of x onto S^{V_i}, whence $x = (x_1, \dots, x_d)$. If $v \in V_j$, then

$$\pi(\Omega(x,v)) = \prod_{i\,:\,i \neq j} \pi_i(x_i) \sum_{\substack{z_j \in S^{V_j} \\ z_j(w)=x_j(w) \text{ for } w \neq v}} \pi_j(z_j).$$

Also, again for $v \in V_j$, if $y \in \Omega(x,v)$, then

$$\pi(y) = \left[\prod_{i\,:\,i \neq j} \pi_i(x_i) \right] \pi_j(y_j).$$

Define for $v \in V_i$ the set $\Omega_i(x,v) := \{z_i \in S^{V_i} : z_i(w) = x_i(w) \text{ for } w \neq v\}$. We have

$$\begin{aligned} P(x,y) &= \sum_{j=1}^{d} \sum_{v \in V_j} \frac{1}{n} \frac{\pi(y)\mathbf{1}\{y \in \Omega(x,v)\}}{\pi(\Omega(x,v))} \\ &= \sum_{j=1}^{d} \frac{n_j}{n} \frac{1}{n_j} \sum_{v \in V_j} \prod_{i\,:\,i \neq j} \mathbf{1}\{y_i = x_i\} \frac{\pi_i(y_i)\mathbf{1}\{y_i \in \Omega_i(x,v)\}}{\pi_i(\Omega_i(x,v))} \\ &= \sum_{j=1}^{d} \frac{n_j}{n} \tilde{P}_j(x,y), \end{aligned}$$

where $\tilde{P}_j$ is the transition matrix of the lift of the Glauber dynamics on S^{V_j} to a chain on S^V. (The lift is defined in (12.20).) The identity (12.21) follows from Corollary 12.12. ■

EXAMPLE 12.15 (Random walk on n-dimensional hypercube). Consider the chain (X_t) on $\Omega := \{-1, 1\}$ which is an i.i.d. sequence of random signs. That is, the transition matrix is

$$P(x,y) = \frac{1}{2} \quad \text{for all } x, y \in \{-1, 1\}. \tag{12.22}$$

Let $I_1(x) = x$, and note that

$$PI_1(x) = \frac{1}{2} + \frac{-1}{2} = 0.$$

Thus there are two eigenfunction: I_1 (with eigenvalue 0) and $\mathbf{1}$, the constant function (with eigenvalue 1).

Consider the lazy random walker on the n-dimensional hypercube, but for convenience write the state space as $\{-1, 1\}^n$. In this state space, the chain moves by selecting a coordinate uniformly at random and refreshing the chosen coordinate with a new random sign, independent of everything else. The transition matrix is exactly (12.19), where each P_j is the two-state transition matrix in (12.22).

By Lemma 12.11, the eigenfunctions are of the form

$$f(x_1, \dots, x_n) = \prod_{j=1}^{n} f_j(x_j)$$

where f_j is either I_1 or $\mathbf{1}$. In other words, for each subset of coordinates $J \subset \{1, 2, \ldots, n\}$,

$$f_J(x_1, \ldots, x_n) := \prod_{j \in J} x_j$$

is an eigenfunction. The corresponding eigenvalue is

$$\lambda_J = \frac{\sum_{i=1}^{n}(1 - \mathbf{1}_{\{i \in J\}})}{n} = \frac{n - |J|}{n}.$$

We take $f_\varnothing(\boldsymbol{x}) := 1$, which is the eigenfunction corresponding to the eigenvalue 1. The eigenfunction $f_{\{1,\ldots,n\}}$ has eigenvalue 0. Each f_J with $|J| = 1$ has corresponding eigenvalue $\lambda_2 = 1 - 1/n$, and consequently $\gamma_\star = 1/n$.

Theorem 12.3 gives

$$t_{\rm mix}(\varepsilon) \leq n\left(-\log\varepsilon + \log(2^n)\right) = n^2\left(\log 2 - n^{-1}\log\varepsilon\right) = O(n^2).$$

Note that this bound is not as good as the bound obtained previously in Section 6.5.2. However, in the next section we will see that careful use of eigenvalues yields a better bound than was obtain in Section 6.5.2.

12.5. An ℓ^2 Bound

For each $p \geq 0$, the $\ell^p(\pi)$ norm on $\mathbb{R}^\Omega$ is defined as

$$\|f\|_p := \left[\sum_{x \in \Omega} |f(x)|^p \pi(x)\right]^{1/p}.$$

An important case is $p = 2$, as $\ell^2(\pi)$ is the inner product space with norm $\|f\|_2 = \sqrt{\langle f, f\rangle_\pi}$.

LEMMA 12.16. *Let P be a reversible transition matrix, with eigenvalues*

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} \geq -1$$

and associated eigenfunctions $\{f_j\}$, orthonormal with respect to $\langle\cdot,\cdot\rangle_\pi$. Then

(i)

$$4\left\|P^t(x,\cdot) - \pi\right\|_{\rm TV}^2 \leq \left\|\frac{P^t(x,\cdot)}{\pi(\cdot)} - 1\right\|_2^2 = \sum_{j=2}^{|\Omega|} f_j(x)^2\lambda_j^{2t}.$$

(ii) *If the chain is transitive, then*

$$4\left\|P^t(x,\cdot) - \pi\right\|_{\rm TV}^2 \leq \left\|\frac{P^t(x,\cdot)}{\pi(\cdot)} - 1\right\|_2^2 = \sum_{j=2}^{|\Omega|} \lambda_j^{2t}.$$

PROOF.

Proof of (i). By Lemma 12.2,

$$\left\|\frac{P^t(x,\cdot)}{\pi(\cdot)} - 1\right\|_2^2 = \left\|\sum_{j=2}^{|\Omega|}\lambda_j^t f_j(x) f_j\right\|_2^2 = \sum_{j=2}^{|\Omega|} f_j(x)^2\lambda_j^{2t}. \tag{12.23}$$

Note that by Proposition 4.2,

$$\left\|P^t(x,\cdot) - \pi\right\|_{\rm TV} = \frac{1}{2}\sum_{y \in \Omega}\left|\frac{P^t(x,y)}{\pi(y)} - 1\right|\pi(y) = \frac{1}{2}\left\|\frac{P^t(x,\cdot)}{\pi(\cdot)} - 1\right\|_1,$$

whence by Exercise 12.5,

$$4\|P^t(x,\cdot)-\pi\|_{\rm TV}^2 = \left\|\frac{P^t(x,\cdot)}{\pi(\cdot)}-1\right\|_1^2 \le \left\|\frac{P^t(x,\cdot)}{\pi(\cdot)}-1\right\|_2^2, \tag{12.24}$$

which with (12.23) establishes (i).

Proof of (ii). Suppose the Markov chain is transitive. Then π is uniform (cf. Proposition 2.16), and the left-hand side of (12.23) does not depend on x. Therefore, for any $x_0 \in \Omega$,

$$\left\|\frac{P^t(x_0,\cdot)}{\pi(\cdot)}-1\right\|_2^2 = \sum_{j=2}^{|\Omega|} f_j(x)^2\lambda_j^{2t}. \tag{12.25}$$

Summing over $x \in X$ on both sides of (12.25),

$$|\Omega|\left\|\frac{P^t(x_0,\cdot)}{\pi(\cdot)}-1\right\|_2^2 = |\Omega|\sum_{j=2}^{|\Omega|}\left[\sum_{x\in\Omega} f_j(x)^2\pi(x)\right]\lambda_j^{2t},$$

where we have multiplied and divided by $\pi(x) = 1/|\Omega|$ on the right-hand side. Since $\|f_j\|_2 = 1$, the inner sum on the right-hand side equals 1, and so

$$\left\|\frac{P^t(x_0,\cdot)}{\pi(\cdot)}-1\right\|_2^2 = \sum_{j=2}^{|\Omega|}\lambda_j^{2t}.$$

Combining with (12.24) establishes (ii). ■

EXAMPLE 12.17. For lazy simple random walk on the hypercube $\{0,1\}^n$, the eigenvalues and eigenfunctions were found in Example 12.15. This chain is transitive, so applying Lemma 12.16 shows that

$$4\|P^t(x,\cdot)-\pi\|_{\rm TV}^2 \le \sum_{k=1}^{n}\left(1-\frac{k}{n}\right)^{2t}\binom{n}{k} \le \sum_{k=1}^{n} e^{-2tk/n}\binom{n}{k} = \left(1+e^{-2t/n}\right)^n - 1. \tag{12.26}$$

Taking $t = (1/2)n\log n + cn$ above shows that

$$4\|P^t(x,\cdot)-\pi\|_{\rm TV}^2 \le \left(1+\frac{1}{n}e^{-2c}\right)^n - 1 \le e^{e^{-2c}} - 1.$$

The right-hand is bounded, for example, by $2e^{-2c}$ provided $c > 1$. Recall that $d(t) := \max_{x\in\Omega}\|P^t(x,\cdot)-\pi\|_{TV}$. The argument in Proposition 7.13 shows that

$$d((1/2)n\log n - cn) \ge 1 - \frac{8}{e^{2c}}\left[1+o(1)\right].$$

Thus we see that in a window of order n around $(1/2)n\log n$, the distance d drops from near one to near zero. This behavior is called ***cutoff*** and is discussed in Chapter 18.

12.6. Time Averages

Suppose that, given a probability distribution π on a finite set Ω and a real-valued function $f : \Omega \to \mathbb{R}$, you want to determine $E_\pi(f) = \sum_{x\in\Omega} f(x)\pi(x)$. If Ω is large or the sum $E_\pi(f)$ is otherwise difficult to compute exactly, then a practical solution may be to estimate $E_\pi(f)$ by averaging f applied to random samples from π.

If you have available an i.i.d. sequence $(X_t)_{t=1}^\infty$ of Ω-valued random elements with common distribution π, then the sequence $(f(X_t))_{t=1}^\infty$ is also i.i.d., each element with expectation $E_\pi(f)$. Because Ω is finite, the variance $\mathrm{Var}(f(X_1))$ of each random variable $f(X_t)$ is finite. The Law of Large Numbers suggests estimating $E_\pi(f)$ by $t^{-1}\sum_{s=1}^t f(X_s)$, and using Chebyshev's inequality, we can give a lower bound on the number of independent samples t needed to ensure that an error of size more than η is made with probability at most ε.

THEOREM 12.18. *Let f be a real-valued function on Ω, and let (X_t) be an i.i.d. sequence of Ω-valued elements, each with distribution π. Then*

$$\mathbf{P}\left\{\left|\frac{1}{t}\sum_{s=1}^{t} f(X_s) - E_\pi(f)\right| > \eta\right\} \le \frac{\mathrm{Var}_\pi(f)}{\eta^2 t}.$$

In particular, if $t \ge \mathrm{Var}_\pi(f)/(\eta^2\varepsilon)$, then the left-hand side is bounded by ε.

The proof is immediate by an application of Chebyshev's inequality to the random variable $t^{-1}\sum_{s=1}^t f(X_s)$, which has variance $t^{-1}\mathrm{Var}_\pi(f)$.

It may be difficult or impossible to get independent exact samples from π. As discussed in Chapter 3, the Markov chain Monte Carlo method is to construct a Markov chain (X_t) for which π is the stationary distribution. In this case, provided that t is a multiple of t_{mix}, the random variable X_t has a distribution close to π. Moreover, X_t and X_{t+s} are approximately independent if s is a multiple of t_{mix}. Thus, in view of Theorem 12.18, one might guess that t should be a multiple of $[\mathrm{Var}_\pi(f)/\eta^2]t_{\mathrm{mix}}$ to ensure that $|t^{-1}\sum_{s=1}^t f(X_s) - E_\pi(f)| < \eta$ with high probability. However, the next theorem shows that after a "burn-in" period of the order of t_{mix}, a multiple of $[\mathrm{Var}_\pi(f)/\eta^2]t_{\mathrm{rel}}$ samples suffices.

THEOREM 12.19. *Let (X_t) be a reversible Markov chain. If $r \ge t_{\mathrm{mix}}(\varepsilon/2)$ and $t \ge [4\,\mathrm{Var}_\pi(f)/(\eta^2\varepsilon)]t_{\mathrm{rel}}$, then for any starting state $x \in \Omega$,*

$$\mathbf{P}_x\left\{\left|\frac{1}{t}\sum_{s=0}^{t-1} f(X_{r+s}) - E_\pi(f)\right| \ge \eta\right\} \le \varepsilon. \tag{12.27}$$

We first prove a lemma needed for the proof of Theorem 12.19.

LEMMA 12.20. *Let (X_t) be a reversible Markov chain and φ an eigenfunction of the transition matrix P with eigenvalue λ and with $\langle\varphi,\varphi\rangle_\pi = 1$. For $\lambda \ne 1$,*

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1}\varphi(X_s)\right)^2\right] \le \frac{2t}{1-\lambda}. \tag{12.28}$$

If f is any real-valued function defined on Ω with $E_\pi(f) = 0$, then

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1} f(X_s)\right)^2\right] \le \frac{2tE_\pi(f^2)}{\gamma}. \tag{12.29}$$

PROOF. For $r < s$,

$$\mathbf{E}_\pi\left[\varphi(X_r)\varphi(X_s)\right] = \mathbf{E}_\pi\left[\mathbf{E}_\pi\left(\varphi(X_r)\varphi(X_s) \mid X_r\right)\right]$$
$$= \mathbf{E}_\pi\left[\varphi(X_r)\,\mathbf{E}_\pi\left(\varphi(X_s) \mid X_r\right)\right] = \mathbf{E}_\pi\left[\varphi(X_r)\,(P^{s-r}\varphi)(X_r)\right].$$

Since φ is an eigenfunction and $E_\pi(\varphi^2) = \langle \varphi, \varphi \rangle_\pi = 1$,

$$\mathbf{E}_\pi\left[\varphi(X_r)\varphi(X_s)\right] = \lambda^{s-r}\mathbf{E}_\pi\left[\varphi(X_r)^2\right] = \lambda^{s-r}E_\pi(\varphi^2) = \lambda^{s-r}.$$

Then by considering separately the diagonal and cross terms when expanding the square,

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1}\varphi(X_s)\right)^2\right] = t + 2\sum_{r=0}^{t-1}\sum_{s=1}^{t-1-r}\lambda^s. \tag{12.30}$$

Evaluating the geometric sum shows that

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1}\varphi(X_s)\right)^2\right] = t + \frac{2t\lambda - 2\lambda(1-\lambda^t)/(1-\lambda)}{1-\lambda}$$
$$= \frac{t(1+\lambda) - 2\lambda g(\lambda)}{1-\lambda},$$

where $g(\lambda) := (1-\lambda^t)/(1-\lambda)$. Note that $g(\lambda) \geq 0$ for $\lambda \in (-1,1)$. When $-1 \leq \lambda \leq 0$, we have $g(\lambda) \leq 1$, whence for $t \geq 2$,

$$t(1+\lambda) - 2\lambda g(\lambda) \leq t(1+\lambda) - t\lambda = t \leq 2t.$$

When $1 > \lambda > 0$, clearly

$$t(1+\lambda) - 2\lambda g(\lambda) \leq t(1+\lambda) \leq 2t.$$

This proves the inequality (12.28).

Let f be a real-valued function on Ω with $E_\pi(f) = 0$. Let $\{f_j\}_{j=1}^{\Omega}$ be the orthonormal eigenfunctions of P of Lemma 12.2. Decompose f as $f = \sum_{j=1}^{|\Omega|} a_j f_j$. By Parseval's Identity, $E_\pi(f^2) = \sum_{j=1}^{|\Omega|} a_j^2$. Observe that $a_1 = \langle f, f_1\rangle_\pi = \langle f, \mathbf{1}\rangle_\pi = E_\pi(f) = 0$.

Defining $G_j := \sum_{s=0}^{t-1} f_j(X_s)$, we can write

$$\sum_{s=0}^{t-1} f(X_s) = \sum_{j=1}^{|\Omega|} a_j G_j.$$

If $r \leq s$ and $j \neq k$, then

$$\begin{aligned}
\mathbf{E}_\pi\left[f_j(X_s)f_k(X_r)\right] &= \mathbf{E}_\pi\left[f_k(X_r)\,\mathbf{E}_\pi(f_j(X_s) \mid X_r)\right] \\
&= \mathbf{E}_\pi\left[f_k(X_r)(P^{s-r}f_j)(X_r)\right] \\
&= \lambda_j^{s-r}\mathbf{E}_\pi\left[f_k(X_r)f_j(X_r)\right] \\
&= \lambda_j^{s-r}E_\pi(f_k f_j) \\
&= 0.
\end{aligned}$$

Consequently, $\mathbf{E}_\pi\left(G_j G_k\right) = 0$ for $j \neq k$. It follows that

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1} f(X_s)\right)^2\right] = \sum_{i=2}^{|\Omega|} a_i^2 \mathbf{E}_\pi\left(G_i^2\right). \tag{12.31}$$

By (12.28), the right-hand side is bounded by

$$\sum_{j=2}^{|\Omega|} \frac{2ta_j^2}{1-\lambda_j} \leq \frac{2tE_\pi(f^2)}{\gamma}.$$

■

PROOF OF THEOREM 12.19. Assume without loss of generality that $E_\pi(f) = 0$; if not, replace f by $f - E_\pi(f)$.

Let μ_r be the optimal coupling of $P^r(x,\cdot)$ with π, which means that

$$\sum_{x \neq y} \mu_r(x,y) = \|P^r(x,\cdot) - \pi\|_{\text{TV}}.$$

We define a process (Y_t, Z_t) as follows: let (Y_0, Z_0) have distribution μ_r. Given (Y_0, Z_0), let (Y_t) and (Z_t) move independently with transition matrix P, until the first time they meet. After they meet, evolve them together according to P. The chain $(Y_t, Z_t)_{t=0}^\infty$ has transition matrix

$$Q((x,y),(z,w)) = \begin{cases} P(x,z) & \text{if } x = y \text{ and } z = w, \\ P(x,z)P(y,w) & \text{if } x \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

The sequences (Y_s) and (Z_s) are each Markov chains with transition matrix P, started with distributions $P^r(x,\cdot)$ and with π, respectively. In particular, $(Y_s)_{s\geq 0}$ has the same distribution as $(X_{r+s})_{s\geq 0}$.

Because the distribution of (Y_0, Z_0) is μ_r,

$$\mathbf{P}\{Y_0 \neq Z_0\} = \|P^r(x,\cdot) - \pi\|_{\text{TV}}. \tag{12.32}$$

Since $(Y_s)_{s\geq 0}$ and $(X_{r+s})_{s\geq 0}$ have the same distribution, we rewrite the probability in (12.27) as

$$\mathbf{P}_x\left\{\left|\frac{1}{t}\sum_{s=0}^{t-1} f(X_{r+s}) - E_\pi(f)\right| > \eta\right\} = \mathbf{P}\left\{\left|\frac{1}{t}\sum_{s=0}^{t-1} f(Y_s) - E_\pi(f)\right| > \eta\right\}.$$

By considering whether or not $Y_0 = Z_0$, this probability is bounded above by

$$\mathbf{P}\{Y_0 \neq Z_0\} + \mathbf{P}\left\{\left|\frac{1}{t}\sum_{s=0}^{t-1} f(Z_s) - E_\pi(f)\right| > \eta\right\}. \tag{12.33}$$

By definition of $t_{\text{mix}}(\varepsilon)$ and the equality (12.32), if $r \geq t_{\text{mix}}(\varepsilon/2)$, then the first term is bounded by $\varepsilon/2$. By Lemma 12.20, the variance of $t^{-1}\sum_{s=0}^{t-1} f(Z_s)$ is bounded by $2\,\text{Var}_\pi(f)/(t\gamma)$. Therefore, Chebyshev's inequality bounds the second term by $\varepsilon/2$, provided that $t \geq [4\,\text{Var}_\pi(f)/(\eta^2\varepsilon)]t_{\text{rel}}$. ■

Exercises

EXERCISE 12.1. Let P be a transition matrix.

(a) Show that all eigenvalues λ of P satisfy $|\lambda| \leq 1$.
Hint: Letting $\|f\|_\infty := \max_{x\in\Omega}|f(x)|$, show that $\|Pf\|_\infty \leq \|f\|_\infty$. Apply this with the eigenfunction φ corresponding to the eigenvalue λ.

(b) Assume P is irreducible. Let $\mathcal{T}(x) = \{t : P^t(x,x) > 0\}$. (Lemma 1.6 shows that $\mathcal{T}(x)$ does not depend on x.) Show that $\mathcal{T}(x) \subset 2\mathbb{Z}$ if and only if -1 is an eigenvalue of P.

(c) Assume P is irreducible, and let ω be an a-th root of unity. Show that $\mathcal{T}(x) \subset a\mathbb{Z}$ if and only if ω is an eigenvalue of P.

EXERCISE 12.2. Let P be irreducible, and suppose that A is a matrix with $0 \leq A(i,j) \leq P(i,j)$ and $A \neq P$. Show that any eigenvalue λ of A satisfies $|\lambda| < 1$.

EXERCISE 12.3. Let $\tilde{P} = (1/2)P + (1/2)I$ be the transition matrix of the lazy version of the chain with transition matrix P. Show that all the eigenvalues of $\tilde{P}$ are non-negative.

EXERCISE 12.4. Show that for a function $f : \Omega \to \mathbb{R}$,

$$\mathrm{Var}_\pi(P^t f) \leq (1-\gamma_\star)^{2t}\,\mathrm{Var}_\pi(f).$$

EXERCISE 12.5. Show that for any $f : \Omega \to \mathbb{R}$, the function $p \mapsto \|f\|_p$ is non-decreasing for $p \geq 1$.

EXERCISE 12.6. Let P be a reversible transition matrix with stationary distribution π. Use Lemma 12.2 to prove that $P^{2t+2}(x,x) \leq P^{2t}(x,x)$.

EXERCISE 12.7. Let P_1 and P_2 by transition matrices on state spaces Ω_1 and Ω_2, respectively. Consider the chain on $\Omega_1 \times \Omega_2$ which moves independently in the first and second coordinates according to P_1 and P_2, respectively. Its transition matrix is the ***tensor product*** $P_1 \otimes P_2$, defined as

$$P_1 \otimes P_2((x,y),(z,w)) = P_1(x,z)P_2(y,w).$$

The tensor product of a function φ on Ω_1 and a function ψ on Ω_2 is the function on $\Omega_1 \times \Omega_2$ defined by $(\varphi \otimes \psi)(x,y) = \varphi(x)\psi(y)$.

Let φ and ψ be eigenfunctions of P_1 and P_2, respectively, with eigenvalues λ and μ. Show that $\varphi \otimes \psi$ is an eigenfunction of $P_1 \otimes P_2$ with eigenvalue $\lambda\mu$.

Notes

Analyzing Markov chains via the eigenvalues of their transition matrices is classical. See Feller (1968, Chapter XVI) or Karlin and Taylor (1981, Chapter 10) (where orthogonal polynomials are used to compute the eigenvalues of certain families of chains). The effectiveness of the ℓ^2 bound was first demonstrated by Diaconis and Shahshahani (1981). Diaconis (1988) uses representation theory to calculate eigenvalues and eigenfunctions for random walks on groups.

Spielman and Teng (1996) show that for any planar graph with n vertices and maximum degree Δ, the relaxation time is at least $c(\Delta)n$, where $c(\Delta)$ is a constant depending on the Δ.

Angel, Peres, and Wilson (2008) analyze the spectral gaps of an interesting family of card shuffles.

For a lazy birth-and-death chain on $\{0,\dots,L\}$, let $\lambda_1,\dots,\lambda_L$ be the eigenvalues of the transition matrix restricted to $\{0,1,\dots,L-1\}$. Then the first hitting time of L starting from 0 has the same distribution as $X_1 + X_2 + \cdots + X_L$, where X_i is geometric with success probability $1-\lambda_i$. A continuous-time version of this was proven in Karlin and McGregor (1959) (see also Keilson (1979) and Fill (2007)). The discrete-time version appears in Diaconis and Fill (1990).

Part II: The Plot Thickens

CHAPTER 13

Eigenfunctions and Comparison of Chains

13.1. Bounds on Spectral Gap via Contractions

In Chapter 5 we used coupling to give a direct bound on the mixing time (cf. Corollary 5.3). We now show that coupling can also be used to obtain bounds on the relaxation time.

THEOREM 13.1 (M. F. Chen (1998)). *Let Ω be a metric space with metric ρ, and let P be the transition matrix of a Markov chain with state space Ω. Suppose there exists a constant $\theta < 1$ such that for each $x, y \in \Omega$ there exists a coupling (X_1, Y_1) of $P(x, \cdot)$ and $P(y, \cdot)$ satisfying*

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq \theta \rho(x, y). \tag{13.1}$$

If $\lambda \neq 1$ is an eigenvalue of P, then $|\lambda| \leq \theta$. In particular, the absolute spectral gap satisfies

$$\gamma_\star \geq 1 - \theta.$$

The ***Lipschitz constant*** of a function $f : \Omega \to \mathbb{R}$ is defined by

$$\mathrm{Lip}(f) := \max_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|f(x) - f(y)|}{\rho(x, y)}.$$

PROOF. For any function f,

$$|Pf(x) - Pf(y)| = |\mathbf{E}_{x,y}\left(f(X_1) - f(Y_1)\right)| \leq \mathbf{E}_{x,y}\left(|f(X_1) - f(Y_1)|\right).$$

By the definition of $\mathrm{Lip}(f)$ and the hypothesis (13.1),

$$|Pf(x) - Pf(y)| \leq \mathrm{Lip}(f)\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq \theta\,\mathrm{Lip}(f)\rho(x, y).$$

This proves that

$$\mathrm{Lip}(Pf) \leq \theta\,\mathrm{Lip}(f).$$

Taking φ to be a non-constant eigenfunction with eigenvalue λ,

$$|\lambda|\,\mathrm{Lip}(\varphi) = \mathrm{Lip}(\lambda\varphi) = \mathrm{Lip}(P\varphi) \leq \theta\,\mathrm{Lip}(\varphi).$$

■

EXAMPLE 13.2 (Metropolis chain for random colorings). Recall the Metropolis chain whose stationary distribution is uniform over all proper q-colorings of a graph, introduced in Example 3.5. At each move this chain picks a vertex v uniformly at random and a color k uniformly at random, then recolors v with k if the resulting coloring is proper.

The proof of Theorem 5.7 constructed, in the case $q > 3\Delta$, a coupling (X_1, Y_1) of $P(x, \cdot)$ with $P(y, \cdot)$ for each pair (x, y) such that

$$\mathbf{E}\left(\rho(X_1, Y_1)\right) \leq \left(1 - \frac{1}{3n\Delta}\right)\rho(x, y).$$

Applying Theorem 13.1 shows that if $q > 3\Delta$, where Δ is the maximum degree of the graph, then

$$\gamma_\star \geq \frac{1}{3n\Delta}.$$

EXAMPLE 13.3. Consider the Glauber dynamics for the hardcore model at fugacity λ, introduced in Section 3.3.4. In the proof of Theorem 5.8, for each pair (x, y), a coupling (X_1, Y_1) of $P(x, \cdot)$ with $P(y, \cdot)$ is constructed which satisfies

$$\mathbf{E}\left(\rho(X_1, Y_1)\right) \leq \left(1 - \frac{1}{n}\left[\frac{1 + \lambda(1 - \Delta)}{1 + \lambda}\right]\right)\rho(x, y).$$

Therefore,

$$\gamma_\star \geq \frac{1}{n}\left[\frac{1 + \lambda(1 - \Delta)}{1 + \lambda}\right].$$

EXAMPLE 13.4. Consider again the lazy random walk on the hypercube $\{0, 1\}^n$, taking the metric to be the Hamming distance $\rho(x, y) = \sum_{i=1}^{d} |x_i - y_i|$.

Let (X_1, Y_1) be the coupling which updates the same coordinate in both chains with the same bit. The distance decreases by one if one among the $\rho(x, y)$ disagreeing coordinates is selected and otherwise remains the same. Thus,

$$\begin{aligned}\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) &\leq \left(1 - \frac{\rho(x, y)}{n}\right)\rho(x, y) + \frac{\rho(x, y)}{n}(\rho(x, y) - 1) \\ &= \left(1 - \frac{1}{n}\right)\rho(x, y).\end{aligned}$$

Applying Theorem 13.1 yields the bound $\gamma_\star \geq n^{-1}$. In Example 12.15 it was shown that $\gamma_\star = n^{-1}$, so the bound of Theorem 13.1 is sharp in this case.

13.2. Wilson's Method for Lower Bounds

A general method due to David Wilson for obtaining a lower bound on mixing time uses an eigenfunction Φ to construct a distinguishing statistic.

THEOREM 13.5 (Wilson's method). *Let (X_t) be an irreducible aperiodic Markov chain with state space Ω and transition matrix P. Let Φ be an eigenfunction of P with eigenvalue λ satisfying $1/2 < \lambda < 1$. Fix $0 < \varepsilon < 1$ and let $R > 0$ satisfy*

$$\mathbf{E}_x\left(|\Phi(X_1) - \Phi(x)|^2\right) \leq R \tag{13.2}$$

for all $x \in \Omega$. Then for any $x \in \Omega$

$$t_{\mathrm{mix}}(\varepsilon) \geq \frac{1}{2\log(1/\lambda)}\left[\log\left(\frac{(1-\lambda)\Phi(x)^2}{2R}\right) + \log\left(\frac{1-\varepsilon}{\varepsilon}\right)\right]. \tag{13.3}$$

At first glance, Theorem 13.5 appears daunting! Yet it gives sharp lower bounds in many important examples. Let's take a closer look and work through an example, before proceeding with the proof.

REMARK 13.6. In applications, ε may not be tiny. For instance, when proving a family of chains has a cutoff, we will need to consider all values $0 < \varepsilon < 1$.

REMARK 13.7. Generally λ will be taken to be the second largest eigenvalue in situations where $\gamma_\star = \gamma = 1 - \lambda$ is small. Under these circumstances a one-term Taylor expansion yields

$$\frac{1}{\log(1/\lambda)} = \frac{1}{\gamma_\star + O(\gamma_\star)^2} = t_{\mathrm{rel}}(1 + O(\gamma_\star)). \tag{13.4}$$

According to Theorems 12.3 and 12.4,

$$\log\left(\frac{1}{2\varepsilon}\right)(t_{\mathrm{rel}} - 1) \leq t_{\mathrm{mix}}(\varepsilon) \leq -\log(\varepsilon\pi_{\min})t_{\mathrm{rel}},$$

where $\pi_{\min} = \min_{x\in\Omega}\pi(x)$. One way to interpret (13.4) is that the denominator of (13.3) gets us up to the relaxation time (ignoring constants, for the moment). The numerator, which depends on the geometry of Φ, determines how much larger a lower bound we can get.

EXAMPLE 13.8. Recall from Example 12.15 that the second-largest eigenvalue of the lazy random walk on the n-dimensional hypercube $\{0,1\}^n$ is $1 - \frac{1}{n}$. The corresponding eigenspace has dimension n, but a convenient representative to take is

$$\Phi(\boldsymbol{x}) = W(\boldsymbol{x}) - \frac{n}{2},$$

where $W(\boldsymbol{x})$ is the Hamming weight (i.e. the number of 1's) in the bitstring $\boldsymbol{x}$. For any bitstring $\boldsymbol{y}$, we have

$$\mathbf{E}_{\boldsymbol{y}}((\Phi(X_1) - \Phi(\boldsymbol{y}))^2) = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2},$$

since the value changes by exactly 1 whenever the walk actually moves. Now apply Theorem 13.5, taking the initial state to be the all-ones vector $\mathbf{1}$ and $R = 1/2$. We get

$$\begin{aligned} t_{\mathrm{mix}}(\varepsilon) &\geq \frac{1}{-2\log(1-n^{-1})}\left\{\log\left[n^{-1}(n/2)^2\right] + \log\left[(1-\varepsilon)/\varepsilon\right]\right\} \\ &= \frac{n}{2}\left[1 + O(n^{-1})\right]\left[\log n + \log[(1-\varepsilon)/\varepsilon] - \log 4\right] \\ &= (1/2)n\log n + (1/2)n[1 + O(n^{-1})]\log[(1-\varepsilon)/\varepsilon] + O(n). \end{aligned}$$

Example 12.17 shows that the leading term $(1/2)n\log n$ is sharp. We obtained a similar lower bound in Proposition 7.13, using the Hamming weight directly as a distinguishing statistic. The major difference between the proof of Proposition 7.13 and the argument given here is that the previous proof used the structure of the hypercube walk to bound the variances. Wilson's method can be seen as a natural (in hindsight!) extension of that argument. What makes Theorem 13.5 widely applicable is that the hypothesis (13.2) is often easily checked and yields good bounds on the variance of the distinguishing statistic $\Phi(X_t)$.

PROOF OF THEOREM 13.5. Since

$$\mathbf{E}(\Phi(X_{t+1})|X_t = z) = \lambda\Phi(z) \tag{13.5}$$

for all $t \geq 0$ and $z \in \Omega$, we have

$$\mathbf{E}_x\Phi(X_t) = \lambda^t\Phi(x) \quad \text{for } t \geq 0 \tag{13.6}$$

by induction. Fix a value t, let $z = X_t$, and define $D_t = \Phi(X_{t+1}) - \Phi(z)$. By (13.5) and (13.2), respectively, we have

$$\mathbf{E}_x(D_t \mid X_t = z) = (\lambda - 1)\Phi(z)$$

and

$$\mathbf{E}_x(D_t^2 \mid X_t = z) \leq R.$$

Hence

$$\begin{aligned} \mathbf{E}_x(\Phi(X_{t+1})^2 \mid X_t = z) &= \mathbf{E}_x((\Phi(z) + D_t)^2 \mid X_t = z) \\ &= \Phi(z)^2 + 2\mathbf{E}_x(D_t\Phi(z) \mid X_t = z) + \mathbf{E}_x(D_t^2 \mid X_t = z) \\ &\leq (2\lambda - 1)\Phi(z)^2 + R. \end{aligned}$$

Averaging over the possible values of $z \in \Omega$ with weights $P^t(x,z) = \mathbf{P}_x\{X_t = z\}$ gives

$$\mathbf{E}_x\Phi(X_{t+1})^2 \leq (2\lambda - 1)\mathbf{E}_x\Phi(X_t)^2 + R.$$

At this point, we could apply this estimate inductively, then sum the resulting geometric series. It is equivalent (and neater) to subtract $R/(2(1-\lambda))$ from both sides, obtaining

$$\mathbf{E}_x\Phi(X_{t+1})^2 - \frac{R}{2(1-\lambda)} \leq (2\lambda - 1)\left(\mathbf{E}_x\Phi(X_t)^2 - \frac{R}{2(1-\lambda)}\right).$$

Iterating the above inequality shows that

$$\mathbf{E}_x\Phi(X_t)^2 - \frac{R}{2(1-\lambda)} \leq (2\lambda - 1)^t\left[\Phi(x)^2 - \frac{R}{2(1-\lambda)}\right].$$

Leaving off the non-positive term $-(2\lambda - 1)^t R/[2(1-\lambda)]$ on the right-hand side above shows that

$$\mathbf{E}_x\Phi(X_t)^2 \leq (2\lambda - 1)^t\Phi(x)^2 + \frac{R}{2(1-\lambda)}. \tag{13.7}$$

Combining (13.6) and (13.7) gives

$$\mathrm{Var}_x\Phi(X_t) \leq \left[(2\lambda - 1)^t - \lambda^{2t}\right]\Phi(x)^2 + \frac{R}{2(1-\lambda)} < \frac{R}{2(1-\lambda)}, \tag{13.8}$$

since $2\lambda - 1 < \lambda^2$ ensures the first term is negative.

Let X_∞ have distribution π and let $t \to \infty$ in (13.6). Theorem 4.9 implies that $\mathbf{E}(\Phi(X_\infty)) = 0$ (as does the orthogonality of eigenfunctions). Similarly, letting $t \to \infty$ in (13.8) gives

$$\mathrm{Var}_x\Phi(X_\infty) \leq \frac{R}{2(1-\lambda)}.$$

Applying Proposition 7.8 with $r^2 = \frac{2(1-\lambda)\lambda^{2t}\Phi(x)^2}{R}$ gives

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \geq \frac{r^2}{4 + r^2} = \frac{(1-\lambda)\lambda^{2t}\Phi(x)^2}{2R + (1-\lambda)\lambda^{2t}\Phi(x)^2}. \tag{13.9}$$

If we take

$$t \leq \frac{1}{2\log(1/\lambda)}\left[\log\left(\frac{(1-\lambda)\Phi(x)^2}{2R}\right) + \log\left(\frac{1-\varepsilon}{\varepsilon}\right)\right]$$

(so that t is at most the right-hand side of (13.3)), then

$$(1-\lambda)\lambda^{2t}\Phi(x)^2 > \frac{\varepsilon}{1-\varepsilon}(2R)$$

and hence the right-hand side of (13.9) is at least ε. ∎

REMARK 13.9. The variance estimate of (13.8) may look crude, but only $O(\lambda^{2t})$ is being discarded. In applications this is generally quite small.

EXAMPLE 13.10 (Product chains). Let P be the transition matrix of a fixed Markov chain with state space Ω, and let Q_n be the transition matrix of the n-dimensional product chain on state space Ω^n, as defined in Section 12.4. At each move, a coordinate is selected at random, and in the chosen coordinate, a transition is made using P. Using Wilson's method, we can derive a lower bound on the mixing time of this family in terms of the parameters of the original chain.

Let $\lambda = \sup_{i\neq 1}\lambda_i$ be the largest non-trivial eigenvalue of P, and let $\gamma = 1-\lambda$. Let $f : \Omega \to \mathbb{C}$ be an eigenfunction of P with eigenvalue λ. By Lemma 12.11, for $1 \leq k \leq n$, the function $\Phi_k : \Omega^n \to \mathbb{C}$ defined by

$$\Phi_k(y_1,\ldots,y_n) = f(y_k)$$

is an eigenfunction of Q_n with eigenvalue

$$\frac{n-1}{n}(1) + \frac{1}{n}(\lambda) = 1 - \frac{\gamma}{n}.$$

Hence $\Phi = \Phi_1 + \cdots + \Phi_n$ is also an eigenfunction with the same eigenvalue.

Let $Y_0, Y_1, Y_2, \ldots$ be a realization of the factor chain, and set

$$R = \sup_{y\in\Omega} \mathbf{E}_y |f(Y_1) - f(y)|^2.$$

Since the product chain moves by choosing a coordinate uniformly and then using P to update that coordinate, the same value of R bounds the corresponding parameter for the product chain Q_n.

Set $m = \max_{y\in\Omega}|f(y)|$. Then applying Theorem 13.5 to the eigenfunction Φ of Q_n tells us that for this product chain,

$$\begin{aligned} t_{\mathrm{mix}}(\varepsilon) &\geq \frac{1}{-2\log\left(1-\frac{\gamma}{n}\right)}\left\{\log\left[\frac{(\gamma/n)n^2m^2}{2R}\right] + \log[(1-\varepsilon)/\varepsilon]\right\} \\ &= \frac{n\log n}{2\gamma} + O(n)\log[(1-\varepsilon)/\varepsilon]. \end{aligned} \tag{13.10}$$

13.3. The Dirichlet Form and the Bottleneck Ratio

13.3.1. The Dirichlet form. Let P be a reversible transition matrix with stationary distribution π. The ***Dirichlet form*** associated to the pair (P,π) is defined for functions f and h on Ω by

$$\mathcal{E}(f,h) := \langle (I-P)f, h\rangle_\pi.$$

LEMMA 13.11. *For a reversible transition matrix P with stationary distribution π, if*

$$\mathcal{E}(f) := \frac{1}{2}\sum_{x,y\in\Omega}[f(x)-f(y)]^2\,\pi(x)P(x,y), \tag{13.11}$$

then $\mathcal{E}(f) = \mathcal{E}(f,f)$.

PROOF. Expanding the square on the right-hand side of (13.11) shows that

$$\mathcal{E}(f) = \frac{1}{2}\sum_{x,y\in\Omega} f(x)^2\pi(x)P(x,y) - \sum_{x,y\in\Omega} f(x)f(y)\pi(x)P(x,y) + \frac{1}{2}\sum_{x,y\in\Omega} f(y)^2\pi(x)P(x,y).$$

By reversibility, $\pi(x)P(x,y) = \pi(y)P(x,y)$, and the first and last terms above are equal to the common value

$$\frac{1}{2}\sum_{x\in\Omega} f(x)^2\pi(x)\sum_{y\in\Omega} P(x,y) = \frac{1}{2}\sum_{x\in\Omega} f(x)^2\pi(x).$$

Therefore,

$$\begin{aligned}\mathcal{E}(f) &= \sum_{x\in\Omega} f(x)^2\pi(x) - \sum_{x\in\Omega} f(x)\left[\sum_{y\in\Omega} f(y)P(x,y)\right]\pi(x)\\ &= \langle f, f\rangle_\pi - \langle f, Pf\rangle_\pi\\ &= \langle f, (I-P)f\rangle_\pi\\ &= \mathcal{E}(f,f).\end{aligned}$$

■

We write $f \perp_\pi g$ to mean $\langle f, g\rangle_\pi = 0$. Let $\mathbf{1}$ denote the function on Ω which is identically 1. Observe that $E_\pi(f) = \langle f, \mathbf{1}\rangle_\pi$, whence $E_\pi(f) = 0$ if and only if $f \perp_\pi \mathbf{1}$.

LEMMA 13.12. *The spectral gap $\gamma = 1 - \lambda_2$ satisfies*

$$\gamma = \min_{\substack{f\in\mathbb{R}^\Omega\\ f\perp_\pi\mathbf{1},\ \|f\|_2=1}} \mathcal{E}(f) = \min_{\substack{f\in\mathbb{R}^\Omega\\ f\perp_\pi\mathbf{1},\ f\not\equiv 0}} \frac{\mathcal{E}(f)}{\|f\|_2^2}. \tag{13.12}$$

REMARK 13.13. Since $\mathcal{E}(f) = \mathcal{E}(f+c)$ for any constant c and $\|f - E_\pi(f)\|_2^2 = \mathrm{Var}_\pi(f)$, if f is a non-constant element of $\mathbb{R}^\Omega$, then

$$\frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)} = \frac{\mathcal{E}(f - E_\pi(f))}{\|f - E_\pi(f)\|_2^2}.$$

Therefore,

$$\gamma = \min_{\substack{f\in\mathbb{R}^\Omega\\ \mathrm{Var}_\pi(f)\neq 0}} \frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)}.$$

PROOF. Let $n = |\Omega|$. As noted in the proof of Lemma 12.2, if $f_1, f_2, \ldots, f_n$ are the eigenfunctions of P associated to the eigenvalues ordered as in (12.6), then $\{f_k\}$ is an orthonormal basis for the inner-product space $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_\pi)$. We can and will always take $f_1 = \mathbf{1}$. Therefore, any function f can be written as $f = \sum_{j=1}^n \langle f, f_j\rangle_\pi f_j$. Computing the $\ell^2(\pi)$ norm of f and using the orthogonality of $\{f_j\}$ shows that

$$\|f\|_2^2 = \langle f, f\rangle_\pi = \sum_{j=1}^{|\Omega|} |\langle f, f_j\rangle_\pi|^2.$$

Accordingly, if $\|f\|_2 = 1$ and $f \perp_\pi \mathbf{1}$ (equivalently, $\langle f, f_1\rangle_\pi = 0$), then $f = \sum_{j=2}^{|\Omega|} a_j f_j$ where $\sum_{j=2}^{|\Omega|} a_j^2 = 1$. Thus,

$$\langle (I-P)f, f\rangle_\pi = \sum_{j=2}^{|\Omega|} a_j^2(1-\lambda_j) \geq 1-\lambda_2,$$

from which follows the first equality in (13.12). To obtain the second equality, for $f \in \mathbb{R}^\Omega$ satisfying $f \perp \mathbf{1}$ and $f \not\equiv 0$, note that $\tilde{f} := f/\|f\|_2$ satisfies $\|\tilde{f}\|_2 = 1$ and $\mathcal{E}(\tilde{f}) = \mathcal{E}f/\|f\|_2^2$. ∎

13.3.2. The bottleneck ratio revisited. We have already met the bottleneck ratio $\Phi_\star$ in Section 7.2, where we established a lower bound on t_{mix} directly in terms of $\Phi_\star$.

The following theorem bounds γ in terms of the bottleneck ratio:

THEOREM 13.14 (Jerrum and Sinclair (1989), Lawler and Sokal (1988)). *Let λ_2 be the second largest eigenvalue of a reversible transition matrix P, and let $\gamma = 1-\lambda_2$. Then*

$$\frac{\Phi_\star^2}{2} \leq \gamma \leq 2\Phi_\star. \tag{13.13}$$

While the lower and upper bounds in Theorem 13.14 look quite different, there exist both examples where the upper bound is the correct order and examples where the lower bound is the correct order. Before proving the theorem, we consider such examples.

EXAMPLE 13.15 (Lazy random walk on the n-dimensional hypercube). Consider the set $S = \{\boldsymbol{x} : x^1 = 0\}$. Then

$$\Phi(S) = 2\sum_{x\in S, y\in S^c} 2^{-n}P(x,y) = 2^{-n+1}2^{n-1}n^{-1}(1/2) = \frac{1}{2n}.$$

Therefore, $\Phi_\star \leq 1/(2n)$. We know that $\gamma = n^{-1}$ (cf. Example 12.15), whence applying Theorem 13.14 shows that

$$\frac{1}{n} \leq 2\Phi_\star \leq \frac{1}{n}.$$

That is, $2\Phi_\star = n^{-1} = \gamma$, showing that for this example, the upper bound in (13.13) is sharp.

EXAMPLE 13.16 (Lazy random walk on the $2n$-cycle). Consider a lazy random walk on a $2n$-cycle. Using the computations in Section 12.3.1 (for the non-lazy chain),

$$\lambda_2 = \frac{\cos(\pi/n)+1}{2} = 1 - \frac{\pi^2}{4n^2} + O(n^{-4}).$$

Therefore, $\gamma = \pi^2/(4n^2) + O(n^{-4})$.

For any set S,

$$\Phi(S) = \frac{|\partial S|\left(\frac{1}{4}\right)\left(\frac{1}{2n}\right)}{\frac{|S|}{2n}}$$

where $\partial S = \{(x,y) : x \in S,\ y \notin S\}$. It is clear that the minimum of $\Phi(S)$ over sets S with $\pi(S) \leq 1/2$ is attained at a segment of length n, whence $\Phi_\star = 1/(2n)$.

The lower bound in (13.13) gives the bound

$$\gamma \geq \frac{1}{8n^2},$$

which is of the correct order.

PROOF OF UPPER BOUND IN THEOREM 13.14. By Lemmas 13.12 and 13.11,

$$\gamma = \min_{\substack{f \not\equiv 0 \\ E_\pi(f)=0}} \frac{\sum_{x,y\in\Omega} \pi(x)P(x,y)\left[f(x)-f(y)\right]^2}{\sum_{x,y\in\Omega} \pi(x)\pi(y)\left[f(x)-f(y)\right]^2}. \tag{13.14}$$

For any S with $\pi(S) \leq 1/2$ define the function f_S by

$$f_S(x) = \begin{cases} -\pi(S^c) & \text{for } x \in S, \\ \pi(S) & \text{for } x \notin S. \end{cases}$$

Since $E_\pi(f_s) = 0$, it follows from (13.14) that

$$\gamma \leq \frac{2Q(S,S^c)}{2\pi(S)\pi(S^c)} \leq \frac{2Q(S,S^c)}{\pi(S)} \leq 2\Phi(S).$$

Since this holds for all S, the upper bound is proved. ■

13.3.3. Proof of lower bound in Theorem 13.14*. We need the following lemma:

LEMMA 13.17. *Given a non-negative function ψ defined on Ω, order Ω so that ψ is non-increasing. If $\pi\{\psi > 0\} \leq 1/2$, then*

$$E_\pi(\psi) \leq \Phi_\star^{-1} \sum_{\substack{x,y\in\Omega \\ x<y}} \left[\psi(x) - \psi(y)\right] Q(x,y).$$

PROOF. Recalling that $\Phi_\star$ is defined as a minimum in (7.6), letting $S = \{x : \psi(x) > t\}$ with $t > 0$ shows that

$$\Phi_\star \leq \frac{Q(S,S^c)}{\pi(S)} = \frac{\sum_{x,y\in\Omega} Q(x,y)\mathbf{1}_{\{\psi(x)>t\geq\psi(y)\}}}{\pi\{\psi > t\}}.$$

Rearranging and noting that $\psi(x) > \psi(y)$ only for $x < y$,

$$\pi\{\psi > t\} \leq \Phi_\star^{-1} \sum_{x<y} Q(x,y)\mathbf{1}_{\{\psi(x)>t\geq\psi(y)\}}.$$

Integrating over t, noting that $\int_0^\infty \mathbf{1}_{\{\psi(x)>t\geq\psi(y)\}}dt = \psi(x) - \psi(y)$, and using Exercise 13.1 shows that

$$E_\pi(\psi) \leq \Phi_\star^{-1} \sum_{x<y} \left[\psi(x) - \psi(y)\right] Q(x,y).$$

■

To complete the proof of the lower bound in Theorem 13.14, let f_2 be an eigenfunction corresponding to the eigenvalue λ_2, so that $Pf_2 = \lambda_2 f_2$. Assume that $\pi\{f_2 > 0\} \leq 1/2$. (If not, use $-f_2$ instead.) Defining $f := \max\{f_2, 0\}$,

$$(I-P)f(x) \leq \gamma f(x) \quad \text{for all } x. \tag{13.15}$$

This is verified separately in the two cases $f(x) = 0$ and $f(x) > 0$. In the former case, (13.15) reduces to $-Pf(x) \leq 0$, which holds because $f(x) \geq 0$. In the case $f(x) > 0$, note that since $f \geq f_2$,

$$(I - P)f(x) \leq (I - P)f_2(x) = (1 - \lambda_2)f_2(x) = \gamma f(x).$$

Because $f \geq 0$,

$$\langle (I - P)f, f\rangle_\pi \leq \gamma \langle f, f\rangle_\pi.$$

Equivalently,

$$\gamma \geq \frac{\langle (I - P)f, f\rangle_\pi}{\langle f, f\rangle_\pi}.$$

Note there is no contradiction to (13.12) because $E_\pi(f) \neq 0$. Applying Lemma 13.17 with $\psi = f^2$ shows that

$$\langle f, f\rangle_\pi^2 \leq \Phi_\star^{-2} \left[\sum_{x<y} \left[f^2(x) - f^2(y)\right] Q(x,y) \right]^2.$$

By the Cauchy-Schwarz inequality,

$$\langle f, f\rangle_\pi^2 \leq \Phi_\star^{-2} \left[\sum_{x<y} \left[f(x) - f(y)\right]^2 Q(x,y) \right] \left[\sum_{x<y} \left[f(x) + f(y)\right]^2 Q(x,y) \right].$$

Using the identity (13.11) of Lemma 13.11 and

$$\left[f(x) + f(y)\right]^2 = 2f^2(x) + 2f^2(y) - \left[f(x) - f(y)\right]^2,$$

we find that

$$\langle f, f\rangle_\pi^2 \leq \Phi_\star^{-2} \langle (I - P)f, f\rangle_\pi \left[2\langle f, f\rangle_\pi - \langle (I - P)f, f\rangle_\pi\right].$$

Let $R := \langle (I - P)f, f\rangle_\pi / \langle f, f\rangle_\pi$ and divide by $\langle f, f\rangle_\pi^2$ to show that

$$\Phi_\star^2 \leq R(2 - R)$$

and

$$1 - \Phi_\star^2 \geq 1 - 2R + R^2 = (1 - R)^2 \geq (1 - \gamma)^2.$$

Finally,

$$\left(1 - \frac{\Phi_\star^2}{2}\right)^2 \geq 1 - \Phi_\star^2 \geq (1 - \gamma)^2,$$

proving that $\gamma \geq \Phi_\star^2/2$, as required.

13.4. Simple Comparison of Markov Chains

If the transition matrix of a chain can be bounded by a constant multiple of the transition matrix for another chain and the stationary distributions of the chains agree, then Lemma 13.12 provides an easy way to compare the spectral gaps. This technique is illustrated by the following example:

EXAMPLE 13.18 (Metropolis and Glauber dynamics for Ising). For a graph with vertex set V with $|V| = n$, let π be the Ising probability measure on $\{-1, 1\}^V$:

$$\pi(\sigma) = Z(\beta)^{-1} \exp\left(\beta \sum_{\substack{v,w \in V \\ v \sim w}} \sigma(v)\sigma(w) \right).$$

(See Section 3.3.5.) The Glauber dynamics chain moves by selecting a vertex v at random and placing a positive spin at v with probability

$$p(\sigma, v) = \frac{e^{\beta S(\sigma,v)}}{e^{\beta S(\sigma,v)} + e^{-\beta S(\sigma,v)}},$$

where $S(\sigma, w) := \sum_{u\,:\,u\sim w} \sigma(u)$. Therefore, if P denotes the transition matrix for the Glauber chain, then for all configurations σ and σ' which differ only at the vertex v, we have

$$P(\sigma, \sigma') = \frac{1}{n} \cdot \frac{e^{\beta\sigma'(v)S(\sigma,v)}}{e^{\beta\sigma'(v)S(\sigma,v)} + e^{-\beta\sigma'(v)S(\sigma,v)}} = \frac{1}{n}\left(\frac{r^2}{1+r^2}\right), \tag{13.16}$$

where $r = e^{\beta\sigma'(v)S(\sigma,v)}$.

We let $\widetilde{P}$ denote the transition matrix for the Metropolis chain using the base chain which selects a vertex v at random and then changes the spin at v. If σ and σ' are two configurations which disagree at the single site v, then

$$\widetilde{P}(\sigma, \sigma') = \frac{1}{n}\left(1 \wedge e^{2\beta\sigma'(v)S(\sigma,v)}\right) = \frac{1}{n}\left(1 \wedge r^2\right). \tag{13.17}$$

(See Section 3.2.)

If $\mathcal{E}$ is the Dirichlet form corresponding to P and $\widetilde{\mathcal{E}}$ is the Dirichlet form corresponding to $\widetilde{P}$, then from (13.16) and (13.17)

$$\frac{1}{2} \le \frac{\mathcal{E}(f)}{\widetilde{\mathcal{E}}(f)} \le 1.$$

Therefore, the gaps are related by

$$\gamma \le \widetilde{\gamma} \le 2\gamma.$$

Example 13.19 (Induced chains). If (X_t) is a Markov chain with transition matrix P, for a non-empty subset $A \subset \Omega$, the ***induced chain on*** A is the chain with state space A and transition matrix

$$P_A(x, y) = \mathbf{P}_x\{X_{\tau_A^+} = y\}$$

for all $x, y \in A$. Intuitively, the induced chain is the original chain, but watched only during the time it spends at states in A.

Theorem 13.20 (Aldous (1999)). *Let (X_t) be a reversible Markov chain on Ω with stationary measure π and spectral gap γ. Let $A \subset \Omega$ be non-empty and let γ_A be the spectral gap for the chain induced on A. Then $\gamma_A \ge \gamma$.*

Proof.

$$\pi(x)P_A(x, y) = \pi(y)P_A(y, x),$$

as is seen by summing over paths, so P_A is reversible with respect to the conditional distribution $\pi_A(B) := \pi(A \cap B)/\pi(A)$. By Lemma 13.12, there exists $\varphi : A \to \mathbb{R}$ with $\langle \varphi, \mathbf{1} \rangle_{\pi_A} = 0$ and

$$\gamma_A = \frac{\mathcal{E}(\varphi)}{\|\varphi\|^2_{\ell^2(\pi_A)}}.$$

Let $\psi : \Omega \to \mathbb{R}$ be the harmonic extension of φ:

$$\psi(x) := \mathbf{E}_x[\varphi(X_{\tau_A})].$$

Observe that for $x \in A$,

$$P\psi(x) = \sum_{y\in\Omega} P(x,y)\psi(y) = \sum_{y\in\Omega} P(x,y)\mathbf{E}_y[\varphi(X_{\tau_A})] = \mathbf{E}_x[\varphi(X_{\tau_A^+})] = P_A\varphi(x).$$

Also, $(I-P)\psi(y) = 0$ for $y \notin A$. Now

$$\begin{aligned}\mathcal{E}(\psi) = \langle (I-P)\psi, \psi\rangle_\pi &= \sum_{x\in A}[(I-P)\psi(x)]\psi(x)\pi(x) \\ &= \sum_{x\in A}[(I-P_A)\varphi(x)]\varphi(x)\pi(x) = \pi(A)\mathcal{E}(\varphi).\end{aligned}$$

Also, writing $\bar{\psi} = \langle \psi, \mathbf{1}\rangle_\pi$, we have

$$\mathrm{Var}_\pi(\psi) \geq \sum_{x\in A}[\varphi(x) - \bar{\psi}]^2\pi(x) \geq \pi(A)\sum_{x\in A}\varphi(x)^2\pi_A(x) = \pi(A)\langle\varphi,\varphi\rangle_{\pi_A}.$$

Thus

$$\gamma \leq \frac{\mathcal{E}(\psi)}{\mathrm{Var}_\pi(\psi)} \leq \frac{\pi(A)\mathcal{E}(\varphi)}{\pi(A)\|\varphi\|^2_{\ell^2(\pi_A)}} = \gamma_A.$$

■

REMARK 13.21. The proof we give above is a bit simpler than Aldous's original proof but follows similar ideas.

The following gives a general comparison between chains when the ratios of both the Dirichlet forms and the stationary distributions can be bounded by constants.

LEMMA 13.22. *Let P and $\tilde{P}$ be reversible transition matrices with stationary distributions π and $\tilde{\pi}$, respectively. If $\tilde{\mathcal{E}}(f) \leq \alpha\mathcal{E}(f)$ for all f, then*

$$\tilde{\gamma} \leq \left[\max_{x\in\Omega}\frac{\pi(x)}{\tilde{\pi}(x)}\right]\alpha\gamma. \tag{13.18}$$

PROOF. Note that $E_\pi(f)$ minimizes $E_\pi(f-\alpha)^2$ among all real values α, and the value attained at the minimum is $\mathrm{Var}_\pi(f)$. Therefore,

$$\mathrm{Var}_\pi(f) \leq E_\pi(f - E_{\tilde{\pi}}(f))^2 = \sum_{x\in\Omega}\left[f(x) - E_{\tilde{\pi}}(f)\right]^2\pi(x).$$

If $c(\pi,\tilde{\pi}) := \max_{x\in\Omega}\pi(x)/\tilde{\pi}(x)$, then the right-hand side above is bounded by

$$c(\pi,\tilde{\pi})\sum_{x\in\Omega}\left[f(x) - E_{\tilde{\pi}}(f)\right]^2\tilde{\pi}(x) = c(\pi,\tilde{\pi})\,\mathrm{Var}_{\tilde{\pi}}(f),$$

whence

$$\frac{1}{\mathrm{Var}_{\tilde{\pi}}(f)} \leq \frac{c(\pi,\tilde{\pi})}{\mathrm{Var}_\pi(f)}. \tag{13.19}$$

By the hypothesis that $\tilde{\mathcal{E}}(f) \leq \alpha\mathcal{E}f$ and (13.19) we see that for any $f \in \mathbb{R}^\Omega$ with $\mathrm{Var}_\pi(f) \neq 0$,

$$\frac{\tilde{\mathcal{E}}(f)}{\mathrm{Var}_{\tilde{\pi}}(f)} \leq \alpha\cdot c(\pi,\tilde{\pi})\cdot\frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)}.$$

By Remark 13.13, taking the minimum over all non-constant $f \in \mathbb{R}^\Omega$ on both sides of the above inequality proves (13.18). ■

13.5. The Path Method

Recall that in Section 5.3.2 we used coupling to show that for lazy simple random walk on the d-dimensional torus $\mathbb{Z}_n^d$ we have $t_{\text{mix}} \leq C_d n^2$. If some edges are removed from the graph (e.g. some subset of the horizontal edges at even heights), then coupling cannot be applied due to the irregular pattern, and the simple comparison techniques of Section 13.4 do not apply, since the sets of allowable transitions do not coincide. In this section, we show how such perturbations of "nice" chains can be studied via comparison. The technique will be exploited later when we study site Glauber dynamics via comparison with block dynamics in Section 15.5 and some further shuffling methods in Chapter 16.

The following theorem—proved in various forms by Jerrum and Sinclair (1989), Diaconis and Stroock (1991), and Quastel (1992), and in the form presented here by Diaconis and Saloff-Coste (1993a)—allows one to compare the behavior of similar reversible chains to achieve bounds on the relaxation time.

For a reversible transition matrix P, define $E = \{(x,y) : P(x,y) > 0\}$. An ***E-path*** from x to y is a sequence $\Gamma = (e_1, e_2, \ldots, e_m)$ of edges in E such that $e_1 = (x, x_1)$, $e_2 = (x_1, x_2)$, $\ldots$, $e_m = (x_{m-1}, y)$ for some vertices $x_1, \ldots, x_{m-1} \in \Omega$. The length of an E-path Γ is denoted by $|\Gamma|$. As usual, $Q(x,y)$ denotes $\pi(x)P(x,y)$.

Let P and $\tilde{P}$ be two reversible transition matrices with stationary distributions π and $\tilde{\pi}$, respectively. Supposing that for each $(x,y) \in \tilde{E}$ there is an E-path from x to y, choose one and denote it by Γ_{xy}. Given such a choice of paths, define the ***congestion ratio*** B by

$$B := \max_{e \in E} \left(\frac{1}{Q(e)} \sum_{\substack{x,y \\ \Gamma_{xy} \ni e}} \tilde{Q}(x,y) |\Gamma_{xy}| \right). \tag{13.20}$$

THEOREM 13.23 (The Comparison Theorem). *Let P and $\tilde{P}$ be reversible transition matrices, with stationary distributions π and $\tilde{\pi}$, respectively. If B is the congestion ratio for a choice of E-paths, as defined in* (13.20), *then*

$$\tilde{\mathcal{E}}(f) \leq B \mathcal{E}(f). \tag{13.21}$$

Consequently,

$$\tilde{\gamma} \leq \left[\max_{x \in \Omega} \frac{\pi(x)}{\tilde{\pi}(x)} \right] B \gamma. \tag{13.22}$$

COROLLARY 13.24. *Let P be a reversible and irreducible transition matrix with stationary distribution π. Suppose Γ_{xy} is a choice of E-path for each x and y, and let*

$$B = \max_{e \in E} \frac{1}{Q(e)} \sum_{\substack{x,y \\ \Gamma_{xy} \ni e}} \pi(x)\pi(y) |\Gamma_{xy}|.$$

Then the spectral gap satisfies $\gamma \geq B^{-1}$.

PROOF. Let $\tilde{P}(x,y) = \pi(y)$, and observe that the stationary measure for $\tilde{P}$ is clearly $\tilde{\pi} = \pi$. For $f \in \mathbb{R}^\Omega$ such that $0 = E_\pi(f) = \langle f, \mathbf{1} \rangle_\pi$,

$$\tilde{\mathcal{E}}(f) = \frac{1}{2} \sum_{x,y \in \Omega} [f(x) - f(y)]^2 \pi(x)\pi(y) = \|f\|_2^2.$$

Applying Theorem 13.23 shows that $\mathcal{E}(f) \geq B^{-1}\|f\|_2^2$. Lemma 13.12 implies that $\gamma \geq B^{-1}$. ■

PROOF OF THEOREM 13.23. For a directed edge $e = (z, w)$, we define $\nabla f(e) := f(w) - f(z)$. Observe that

$$2\tilde{\mathcal{E}}(f) = \sum_{(x,y)\in\tilde{E}} \tilde{Q}(x,y)[f(x) - f(y)]^2 = \sum_{x,y} \tilde{Q}(x,y) \left[\sum_{e\in\Gamma_{x,y}} \nabla f(e)\right]^2.$$

Applying the Cauchy-Schwarz inequality yields

$$2\tilde{\mathcal{E}}(f) \leq \sum_{x,y} \tilde{Q}(x,y)|\Gamma_{xy}| \sum_{e\in\Gamma_{x,y}} [\nabla f(e)]^2 = \sum_{e\in E} \left[\sum_{\Gamma_{xy}\ni e} \tilde{Q}(x,y)|\Gamma_{xy}|\right] [\nabla f(e)]^2.$$

By the definition of the congestion ratio, the right-hand side is bounded above by

$$\sum_{(z,w)\in E} BQ(z,w)[f(w) - f(z)]^2 = 2B\mathcal{E}(f),$$

completing the proof of (13.21).

The inequality (13.22) follows from Lemma 13.22. ■

EXAMPLE 13.25 (Comparison for simple random walks on graphs). If two graphs have the same vertex set but different edge sets E and $\tilde{E}$, then

$$Q(x,y) = \frac{1}{2|E|}\mathbf{1}_{(x,y)\in E} \quad \text{and} \quad \tilde{Q}(x,y) = \frac{1}{2|\tilde{E}|}\mathbf{1}_{(x,y)\in\tilde{E}}.$$

Therefore, the congestion ratio is simply

$$B = \left(\max_{e\in E} \sum_{\Gamma_{xy}\ni e} |\Gamma_{xy}|\right) \frac{|E|}{|\tilde{E}|}.$$

In our motivating example, we only removed horizontal edges at even heights from the torus. Since all odd-height edges remain, we can take $|\Gamma_{xy}| \leq 3$ since we can traverse any missing edge in the torus by moving upwards, then across the edge of odd height, and then downwards. The horizontal edge in this path would then be used by at most 3 paths Γ (including the edge itself). Since we removed at most one quarter of the edges, $B \leq 12$.

Thus the relaxation time for the perturbed torus also satisfies $t_{\text{rel}} = O(n^2)$.

13.5.1. Averaging over paths. In Theorem 13.23, for each $e = (x, y) \in \tilde{E}$ we select a single path Γ_{xy} from x to y using edges in E. Generally there will be many paths between x and y using edges from E, and it is often possible to reduce the worst-case bottlenecking specifying by a measure ν_{xy} on the set $\mathcal{P}_{xy}$ of paths from x to y. One can think of this measure as describing how to select a random path between x and y.

In this case, the congestion ratio is given by

$$B := \max_{e\in E} \left(\frac{1}{Q(e)} \sum_{(x,y)\in\tilde{E}} \tilde{Q}(x,y) \sum_{\Gamma: e\in\Gamma\in\mathcal{P}_{xy}} \nu_{xy}(\Gamma)|\Gamma|\right). \tag{13.23}$$

COROLLARY 13.26. *Let P and $\tilde{P}$ be two reversible transition matrices with stationary distributions π and $\tilde{\pi}$, respectively. If B is the congestion ratio for a choice of randomized E-paths, as defined in* (13.23), *then*

$$\tilde{\gamma} \leq \left[\max_{x\in\Omega} \frac{\pi(x)}{\tilde{\pi}(x)}\right] B\gamma. \tag{13.24}$$

The proof of Corollary 13.26 is exactly parallel to that of Theorem 13.23. Exercise 13.3 asks you to fill in the details.

13.5.2. Comparison of random walks on groups. When the two Markov chains that we are attempting to compare are both random walks on the same group G, it is enough to write the support of the increments of one walk in terms of the support of the increments of the other. Then symmetry can be used to get an evenly-distributed collection of paths.

To fix notation, let μ and $\tilde{\mu}$ be the increment measures of two irreducible and reversible random walks on a finite group G. Let S and $\tilde{S}$ be the support sets of μ and $\tilde{\mu}$, respectively, and, for each $a \in \tilde{S}$, fix an expansion $a = s_1 \dots s_k$, where $s_i \in S$ for $1 \leq i \leq k$. Write $N(s,a)$ for the number of times $s \in S$ appears in the expansion of $a \in \tilde{S}$, and let $|a| = \sum_{s\in S} N(s,a)$ be the total number of factors in the expansion of a.

In this case the appropriate congestion ratio is

$$B := \max_{s\in S} \frac{1}{\mu(s)} \sum_{a\in\tilde{S}} \tilde{\mu}(a)N(s,a)\,|a|. \tag{13.25}$$

COROLLARY 13.27. *Let μ and $\tilde{\mu}$ be the increment measures of two irreducible and reversible random walks on a finite group G. Let γ and $\tilde{\gamma}$ be their spectral gaps, respectively.*

Then

$$\tilde{\gamma} \leq B\gamma, \tag{13.26}$$

where B is the congestion ratio defined in (13.25).

PROOF. Let P and $\tilde{P}$ be the transition matrices of the random walks on G with increment measures μ and $\tilde{\mu}$, respectively. Let $E = \{(g,h) | P(g,h) > 0\}$. For $e = (g,h) \in E$, we have

$$Q(e) = Q(g,h) = \frac{P(g,h)}{|G|} = \frac{\mu(hg^{-1})}{|G|}.$$

(Recall that the uniform distribution is stationary for every random walk on G.) Define $\tilde{E}$ and $\tilde{Q}$ in a parallel way.

To obtain a path corresponding to an arbitrary edge $(b,c) \in \tilde{E}$, write $c = ab$ where $a \in \tilde{S}$ has generator expansion $s_1 \dots s_k$. Then

$$c = s_1 \dots s_k b$$

determines a path Γ_{bc} from b to c using only edges in E.

We now estimate the congestion ratio

$$\max_{e\in E} \left(\frac{1}{Q(e)} \sum_{\substack{g,h \\ \Gamma_{gh} \ni e}} \tilde{Q}(g,h)|\Gamma_{gh}| \right). \tag{13.27}$$

For how many pairs $\{g,h\} \in \tilde{E}$ does a specific $e \in E$ appear in Γ_{gh}? Let $s \in S$ be the generator corresponding to e, that is, $e = \{b, sb\}$ for some $b \in G$. For every occurrence of an edge $\{c, sc\}$ using s in the generator path for some $a \in \tilde{S}$, the edge e appears in the path for $\{c^{-1}b, ac^{-1}b\} \in \tilde{E}$.

Hence the congestion ratio simplifies to

$$B = \max_{e \in E} \left(\frac{|G|}{P(e)} \sum_{\substack{g,h \\ \Gamma_{gh} \ni e}} \frac{\tilde{P}(g,h)}{|G|} |\Gamma_{gh}| \right) = \max_{s \in S} \frac{1}{\mu(s)} \sum_{a \in \tilde{S}} N(s,a)|a|\tilde{\mu}(a).$$

Applying Theorem 13.23 completes the proof. ■

REMARK 13.28. The generalization to randomized paths goes through in the group case just as it does for general reversible chains (Corollary 13.26). We must now for each generator $a \in \tilde{S}$ specify a measure ν_a on the set $\mathcal{P}_a = \{(s_1, \ldots, s_k) : s_1 \cdots s_k = a\}$ of expansions of a in terms of elements of S. If we let $|\Gamma|$ be the number of elements in an expansion $\Gamma = (s_1, \ldots, s_k)$ and $N(a, \Gamma)$ be the number of times a appears in Γ, then the appropriate congestion ratio is

$$B := \max_{s \in S} \frac{1}{\mu(s)} \sum_{a \in \tilde{S}} \tilde{\mu}(a) \sum_{\Gamma \in \mathcal{P}_a} \nu_a(\Gamma) N(s,\Gamma)\, |\Gamma|. \tag{13.28}$$

Exercise 13.4 asks you to fill in the details.

Using randomized paths can be useful, for example, when the generating set S of the "new" walk is much larger than the generating set $\tilde{S}$ of the already-understood walk; in such a case averaging over paths can spread the bottlenecking over all generators, rather than just a few.

13.6. Expander Graphs*

When a graph has a narrow bottleneck, the corresponding random walk must mix slowly. How efficiently can a family of graphs avoid bottlenecks? What properties does such an optimal family enjoy?

A family $\{G_n\}$ of graphs is defined to be a (d, α)-***expander family*** if the following three conditions hold for all n:

(i) $\lim_{n \to \infty} |V(G_n)| = \infty$.
(ii) G_n is d-regular.
(iii) The bottleneck ratio of simple random walk on G_n satisfies $\Phi_\star(G_n) \geq \alpha$.

PROPOSITION 13.29. *When $\{G_n\}$ is a (d,α)-expander family, the lazy random walks on $\{G_n\}$ satisfy $t_{\text{mix}}(G_n) = O(\log |V(G_n)|)$.*

PROOF. Theorem 13.14 implies that for all G_n the spectral gap for the simple random walk satisfies $\gamma \geq \alpha^2/2$. Since each G_n is regular, the stationary distribution of the lazy random walk is uniform, and Theorem 12.3 tells us that for the lazy walk $t_{\text{mix}}(G_n) = O(\log |V(G_n)|)$. ■

REMARK 13.30. Given the diameter lower bound of Section 7.1.2, Proposition 13.29 says that expander families exhibit the fastest possible mixing (up to constant factors) for families of graphs of bounded degree.

It is not at all clear from the definition that families of expanders exist. Below we construct a family of 3-regular expander graphs. This is a version of the first construction of an expander family, due to Pinsker (1973). Our initial construction allows multiple edges; we then describe modifications that yield 3-regular simple graphs.

Let $V(G_n) = \{a_1, \ldots, a_n, b_1, \ldots, b_n\}$. Choose permutations $\sigma_1, \sigma_2 \in \mathcal{S}_n$ uniformly at random and independent of each other, and set

$$E(G_n) = \{(a_i, b_i), (a_i, b_{\sigma_1(i)}), (a_i, b_{\sigma_2(i)}) \,:\, 1 \leq i \leq n\}. \tag{13.29}$$

PROPOSITION 13.31. *For the family $\{G_n\}$ of random multigraphs described in* (13.29),

$$\lim_{n\to\infty} \mathbf{P}\{\Phi_\star(G_n) > 0.01\} = 1.$$

PROOF. Assume that $\delta < .03$. We first show that *with probability tending to* 1 *as $n \to \infty$, every subset of A of size $k \leq n/2$ has more than $(1+\delta)k$ neighbors.* Note that every edge in G_n connects a vertex in $A = \{a_1, \ldots, a_n\}$ to a vertex in $B = \{b_1, \ldots, b_n\}$ (that is, G_n is bipartite).

Let $S \subset A$ be a set of size $k \leq n/2$, and let $N(S)$ be the set of neighbors of S. We wish to bound the probability that $|N(S)| \leq (1+\delta)k$. Since (a_i, b_i) is an edge for any $1 \leq i \leq n$, we get immediately that $|N(S)| \geq k$. We can upper bound the probability that $N(S)$ is small by first enumerating the possibilities for the set of δk "surplus" vertices allowed in $N(S)$, then making sure both $\sigma_1(S)$ and $\sigma_2(S)$ fall within the specified set. This argument gives

$$\mathbf{P}\left\{|N(S)| \leq (1+\delta)k\right\} \leq \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{k}^2}{\binom{n}{k}^2},$$

so

$$\mathbf{P}\left\{\text{for some } S,\ |S| \leq n/2 \text{ and } |N(S)| \leq (1+\delta)k\right\} \leq \sum_{k=1}^{n/2} \binom{n}{k} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}^2}.$$

Exercise 13.5 asks you to show that this summation tends to 0 for $\delta < .03$.

We finish by checking that *if every subset of A of size $k \leq n/2$ has more than $(1+\delta)k$ neighbors, then $\Phi_\star > \delta/2$.* Why? For $S \subset V$ with $|S| \leq n$, let

$$A' = S \cap A \qquad \text{and} \qquad B' = S \cap B.$$

Without loss of generality we may assume $|A'| \geq |B'|$. If $|A'| \leq n/2$, then by hypothesis A' has more than $(\delta/2)|S|$ neighbors in $B - B'$: all those edges connect elements of S to elements of S^c. If $|A'| \geq n/2$, let $A'' \subseteq A'$ be an arbitrary subset of size $\lceil n/2 \rceil$. Again, A'' must have more than $(\delta/2)|S|$ neighbors in $B - B'$, and all the corresponding edges connect S and S^c.

Taking $\delta = .02$ completes the proof. ■

COROLLARY 13.32. *There exists a family of* $(3, 0.001)$*-expanders.*

PROOF. We claim first that we can find a family of (deterministic) 3-regular multigraphs $\{G_n\}$ such that each has at most 100 double edges, no triple edges, and bottleneck ratio at least 0.01. Why? Proposition 13.31 guarantees that asymptotically almost every random graph in the model of (13.29) has the bottleneck ratio property. Since a triple edge implies that the two vertices involved are disconnected from the rest of the graph, no graph with the bottleneck property has triple



FIGURE 13.1. Modifying a 3-regular multigraph to get a 3-regular graph.

edges. Furthermore, the expectation and variance of the number of double edges produced by each pair of the three permutations $\{\text{id}, \sigma_1, \sigma_2\}$ are both 1 (as can be easily checked using indicators), so the probability of a total of more than 100 double edges is much less than 1. For large enough n, this event must have non-trivial intersection with the event that the bottleneck ratio is at least 0.01. We select one graph in that intersection for each sufficiently large n.

We still must repair the double edges. Subdivide each one with a vertex; then connect the two added vertices with an edge (as shown in Figure 13.1). Call the resulting graphs $\{\widetilde{G_n}\}$. These modifications will have a negligible effect on the bottleneck ratio for sufficiently large n. ■

REMARK 13.33. In fact, as n tends to ∞, the probability that G_n is a simple graph tends to $1/e^3$—see Riordan (1944). Verifying this fact (which we will not do here) also suffices to demonstrate the existence of an expander family.

Exercises

EXERCISE 13.1. Let Y be a non-negative random variable. Show that

$$\mathbf{E}(Y) = \int_0^\infty \mathbf{P}\{Y > t\} dt.$$

Hint: Write $Y = \int_0^\infty \mathbf{1}_{\{Y>t\}} dt$.

EXERCISE 13.2. Show that for lazy simple random walk on the box $\{1, \ldots, n\}^d$, the parameter $\gamma_\star$ satisfies $\gamma_\star^{-1} = O(n^2)$.

EXERCISE 13.3. Prove Corollary 13.26. *Hint*: follow the outline of the proof of Theorem 13.23.

EXERCISE 13.4. Prove that the statement of Corollary 13.27 remains true in the situation outlined in Remark 13.28.

EXERCISE 13.5. To complete the proof of Proposition 13.31, prove that for $\delta < 0.03$

$$\lim_{n\to\infty} \sum_{k=1}^{n/2} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} = 0.$$

Notes

Wilson's method first appeared in Wilson (2004a). Wilson (2003) extended his lower bound to complex eigenvalues. See Mossel et al. (2004) for another variant.

The connection between the spectral gap of the Laplace-Beltrami operator on Riemannian manifolds and an isoperimetric constant is due to Cheeger (1970); hence the bottleneck ratio is often called the ***Cheeger constant***. The relationship

between the bottleneck ratio and the spectral gap for random walks on graphs was observed by Alon and Milman (1985) and Alon (1986). The method of canonical paths for bounding relaxation time was introduced in Sinclair and Jerrum (1989) and Lawler and Sokal (1988), then further developed in Diaconis and Stroock (1991) and Sinclair (1992).

The bottleneck constant is also sometimes called ***conductance***, especially in the computer science literature. We avoid this term, because it clashes with our use of "conductance" for electrical networks in Chapter 9.

The Comparison Theorem is an extension of the method of canonical paths. A special case appeared in Quastel (1992); the form we give here is from Diaconis and Saloff-Coste (1993a) and Diaconis and Saloff-Coste (1993b). See also Madras and Randall (1996), Randall and Tetali (2000), and Dyer, Goldberg, Jerrum, and Martin (2006). Considering random paths, rather than a "canonical" path between each pair of states, is sometimes called the method of ***multicommodity flows***. We avoid this term because it clashes (partially) with our use of "flow" in Chapter 9. Here a probability measure on paths for x to y clearly determines a unit flow from x to y; however, a flow by itself does not contain enough information to determine the congestion ratio of (13.23).

Pinsker (1973) showed that random regular graphs are expanders. Expander graphs are used extensively in computer science and communications networks. See Sarnak (2004) for a brief exposition and Hoory, Linial, and Wigderson (2006) or Lubotzky (1994) for a full discussion, including many deterministic constructions.

Complements. Theorem 13.1 can be combined with Theorem 12.3 to get a bound on mixing time when there is a coupling which contracts, in the reversible case: If for each pair of states x, y, there exists a coupling (X_1, Y_1) of $P(x, \cdot)$ and $P(y, \cdot)$ satisfying

$$\mathbf{E}_{x,y}(\rho(X_1, Y_1)) \leq \theta \rho(x, y),$$

then

$$t_{\text{mix}}(\varepsilon) \leq \frac{-\log(\varepsilon) - \log(\pi_{\min})}{1 - \theta}. \tag{13.30}$$

Compare with Corollary 14.7, which bounds mixing time directly from a contractive coupling. Since $\pi_{\min}\text{diam} \leq \pi_{\min}|\Omega| \leq 1$, it follows that $-\log(\pi_{\min}) \geq \log(\text{diam})$ and the bound in (13.30) is never better than the bound given by Corollary 14.7. In fact, (13.30) can be much worse. For example, for the hypercube, $\pi_{\min}^{-1} = 2^d$, while the diameter is d.

CHAPTER 14

The Transportation Metric and Path Coupling

Let P be a transition matrix on a metric space (Ω, ρ), where the metric ρ satisfies $\rho(x, y) \geq \mathbf{1}\{x \neq y\}$. Suppose, for all states x and y, there exists a coupling (X_1, Y_1) of $P(x, \cdot)$ with $P(y, \cdot)$ that contracts ρ on average, i.e., which satisfies

$$\mathbf{E}_{x,y}\rho(X_1, Y_1) \leq e^{-\alpha}\rho(x, y) \tag{14.1}$$

for some $\alpha > 0$. The ***diameter*** of Ω is defined to be $\text{diam}(\Omega) := \max_{x,y\in\Omega} \rho(x, y)$. By iterating (14.1), we have

$$\mathbf{E}_{x,y}\rho(X_t, Y_t) \leq e^{-\alpha t}\text{diam}(\Omega).$$

We conclude that

$$\begin{aligned}\left\|P^t(x, \cdot) - P^t(y, \cdot)\right\|_{\text{TV}} \leq \mathbf{P}_{x,y}\{X_t \neq Y_t\} &= \mathbf{P}_{x,y}\{\rho(X_t, Y_t) \geq 1\} \\ &\leq \mathbf{E}_{x,y}\rho(X_t, Y_t) \leq \text{diam}(\Omega)e^{-\alpha t},\end{aligned}$$

whence

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{1}{\alpha}\left[\log(\text{diam}(\Omega)) + \log(1/\varepsilon)\right]\right\rceil.$$

This is the method used in Theorem 5.7 to bound the mixing time of the Metropolis chain for proper colorings and also used in Theorem 5.8 for the hardcore chain.

Path coupling is a technique that simplifies the construction of couplings satisfying (14.1), when ρ is a *path metric*, defined below. While the argument just given requires verification of (14.1) *for all* pairs $x, y \in \Omega$, the path-coupling technique shows that it is enough to construct couplings satisfying (14.1) only for neighboring pairs.

14.1. The Transportation Metric

Recall that a coupling of probability distributions μ and ν is a pair (X, Y) of random variables defined on a single probability space such that X has distribution μ and Y has distribution ν.

For a given distance ρ defined on the state space Ω, the ***transportation metric*** between two distributions on Ω is defined by

$$\rho_K(\mu, \nu) := \inf\{\mathbf{E}(\rho(X, Y)) \,:\, (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{14.2}$$

By Proposition 4.7, if $\rho(x, y) = \mathbf{1}_{\{x\neq y\}}$, then $\rho_K(\mu, \nu) = \|\mu - \nu\|_{\text{TV}}$.

REMARK 14.1. It is sometimes convenient to describe couplings using probability distributions on the product space $\Omega\times\Omega$, instead of random variables. When q is a probability distribution on $\Omega\times\Omega$, its ***projection onto the first coordinate*** is the probability distribution on Ω equal to

$$q(\cdot \times \Omega) = \sum_{y\in\Omega} q(\cdot, y).$$

Likewise, its ***projection onto the second coordinate*** is the distribution $q(\Omega \times \cdot)$.

Given a coupling (X, Y) of μ and ν as defined above, the distribution of (X, Y) on $\Omega \times \Omega$ has projections μ and ν on the first and second coordinates, respectively. Conversely, given a probability distribution q on $\Omega \times \Omega$ with projections μ and ν, the identity function on the probability space $(\Omega \times \Omega, q)$ is a coupling of μ and ν.

Consequently, since $\mathbf{E}(\rho(X, Y)) = \sum_{(x,y)\in\Omega\times\Omega} \rho(x, y)q(x, y)$ when (X, Y) has distribution q, the transportation metric can also be written as

$$\rho_K(\mu, \nu) = \inf \left\{ \sum_{(x,y)\in\Omega\times\Omega} \rho(x, y)q(x, y) \ : \ q(\cdot \times \Omega) = \mu, \ q(\Omega \times \cdot) = \nu \right\}. \tag{14.3}$$

REMARK 14.2. The set of probability distributions on $\Omega \times \Omega$ can be identified with the ($|\Omega|^2$-1)-dimensional simplex, which is a compact subset of $\mathbb{R}^{|\Omega|^2}$. The set of distributions on $\Omega \times \Omega$ which project on the first coordinate to μ and project on the second coordinate to ν is a closed subset of this simplex and hence is compact. The function

$$q \mapsto \sum_{(x,y)\in\Omega\times\Omega} \rho(x, y)q(x, y)$$

is continuous on this set. Hence there is a $q_\star$ such that

$$\sum_{(x,y)\in\Omega\times\Omega} \rho(x, y)q_\star(x, y) = \rho_K(\mu, \nu).$$

Such a $q_\star$ is called an ***optimal coupling*** of μ and ν. Equivalently, there is a pair of random variables $(X_\star, Y_\star)$, also called an optimal coupling, such that

$$\mathbf{E}(\rho(X_\star, Y_\star)) = \rho_K(\mu, \nu).$$

LEMMA 14.3. *The function ρ_K defined in (14.2) is a metric on the space of probability distributions on Ω.*

PROOF. We check the triangle inequality and leave the verification of the other two conditions to the reader.

Let μ, ν and η be probability distributions on Ω. Let p be a probability distribution on $\Omega \times \Omega$ which is a coupling of μ and ν, and let q be a probability distribution on $\Omega \times \Omega$ which is a coupling of ν and η. Define the probability distribution r on $\Omega \times \Omega \times \Omega$ by

$$r(x, y, z) := \frac{p(x, y)q(y, z)}{\nu(y)}. \tag{14.4}$$

(See Remark 14.4 for the motivation of this definition.) Note that the projection of r onto its first two coordinates is p, and the projection of r onto its last two coordinates is q. The projection of r onto the first and last coordinates is a coupling of μ and η.

Assume now that p is an optimal coupling of μ and ν. (See Remark 14.2.) Likewise, suppose that q is an optimal coupling of ν and η.

Let (X, Y, Z) be a random vector with probability distribution r. Since ρ is a metric,

$$\rho(X, Z) \leq \rho(X, Y) + \rho(Y, Z).$$

Taking expectation, because (X, Y) is an optimal coupling of μ and ν and (Y, Z) is an optimal coupling of ν and η,

$$\mathbf{E}(\rho(X, Z)) \leq \mathbf{E}(\rho(X, Y)) + \mathbf{E}(\rho(Y, Z)) = \rho_K(\mu, \nu) + \rho_K(\nu, \eta).$$

Since (X, Z) is a coupling (although not necessarily optimal) of μ and η, we conclude that

$$\rho_K(\mu, \eta) \leq \rho_K(\mu, \nu) + \rho_K(\nu, \eta).$$

■

The transportation metric ρ_K extends the metric ρ on Ω to a metric on the space of probability distributions on Ω. In particular, if δ_x denotes the probability distribution which puts unit mass on x, then $\rho_K(\delta_x, \delta_y) = \rho(x, y)$.

REMARK 14.4. The probability distribution r defined in (14.4) can be thought of as three steps of a time-inhomogeneous Markov chain. The first state X is generated according to μ. Given $X = x$, the second state Y is generated according to $p(x, \cdot)/\mu(x)$, and given $Y = y$, the third state Z is generated according to $q(y, \cdot)/\nu(y)$. Thus,

$$\mathbf{P}\{X = x, Y = y, Z = z\} = \mu(x)\frac{p(x, y)}{\mu(x)}\frac{q(y, z)}{\nu(y)} = r(x, y, z).$$

14.2. Path Coupling

Suppose that the state space Ω of a Markov chain (X_t) is the vertex set of a connected graph $G = (\Omega, E_0)$ and ℓ is a length function defined on E_0. That is, ℓ assigns length $\ell(x, y)$ to each edge $\{x, y\} \in E_0$. We assume that $\ell(x, y) \geq 1$ for all edges $\{x, y\}$.

REMARK 14.5. This graph structure may be different from the structure inherited from the permissible transitions of the Markov chain (X_t).

Define a ***path*** in Ω from x to y to be a sequence of states $\xi = (x_0, x_1, \ldots, x_r)$ such that $x_0 = x$ and $x_r = y$ and such that $\{x_{i-1}, x_i\}$ is an edge for $i = 1, \ldots, r$. The ***length*** of the path is defined to be $\sum_{i=1}^{r} \ell(x_{i-1}, x_i)$. The ***path metric*** on Ω is defined by

$$\rho(x, y) = \min\{\text{length of } \xi : \xi \text{ a path from } x \text{ to } y\}. \tag{14.5}$$

Since we have assumed that $\ell(x, y) \geq 1$, it follows that $\rho(x, y) \geq \mathbf{1}\{x \neq y\}$, whence for any pair (X, Y),

$$\mathbf{P}\{X \neq Y\} = \mathbf{E}\left(\mathbf{1}_{\{X \neq Y\}}\right) \leq \mathbf{E}\rho(X, Y). \tag{14.6}$$

Minimizing over all couplings (X, Y) of μ and ν shows that

$$\rho_{TV}(\mu, \nu) \leq \rho_K(\mu, \nu). \tag{14.7}$$

While Bubley and Dyer (1997) discovered the following theorem and applied it to mixing, the key idea is the application of the triangle inequality for the transportation metric, which goes back to Kantorovich (1942).

THEOREM 14.6 (Bubley and Dyer (1997)). *Suppose the state space Ω of a Markov chain is the vertex set of a graph with length function ℓ defined on edges. Let ρ be the corresponding path metric defined in* (14.5). *Suppose that for each edge $\{x, y\}$ there exists a coupling (X_1, Y_1) of the distributions $P(x, \cdot)$ and $P(y, \cdot)$ such that*

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq \rho(x, y)e^{-\alpha} = \ell(x, y)e^{-\alpha}. \tag{14.8}$$

Then for any two probability measures μ and ν on Ω,

$$\rho_K(\mu P, \nu P) \leq e^{-\alpha}\rho_K(\mu, \nu). \tag{14.9}$$

Recall that $d(t) = \max_{x\in\Omega} \|P^t(x,\cdot) - \pi\|_{TV}$ and $\mathrm{diam}(\Omega) = \max_{x,y\in\Omega} \rho(x,y)$.

COROLLARY 14.7. *Suppose that the hypotheses of Theorem 14.6 hold. Then*

$$d(t) \le e^{-\alpha t}\mathrm{diam}(\Omega),$$

and consequently

$$t_{\mathrm{mix}}(\varepsilon) \le \left\lceil \frac{-\log(\varepsilon) + \log(\mathrm{diam}(\Omega))}{\alpha} \right\rceil .$$

PROOF. By iterating (14.9), it follows that

$$\rho_K(\mu P^t, \nu P^t) \le e^{-\alpha t}\rho_K(\mu,\nu) \le e^{-\alpha t} \max_{x,y} \rho(x,y). \tag{14.10}$$

Applying (14.7) and setting $\mu = \delta_x$ and $\nu = \pi$ shows that

$$\left\|P^t(x,\cdot) - \pi\right\|_{TV} \le e^{-\alpha t}\mathrm{diam}(\Omega). \tag{14.11}$$

■

PROOF OF THEOREM 14.6. We begin by showing that for arbitrary (not necessarily neighboring) $x, y \in \Omega$,

$$\rho_K(P(x,\cdot), P(y,\cdot)) \le e^{-\alpha}\rho(x,y). \tag{14.12}$$

Fix $x, y \in \Omega$, and let $(x_0, x_1, \dots, x_r)$ be a path achieving the minimum in (14.5). By the triangle inequality for ρ_K,

$$\rho_K(P(x,\cdot), P(y,\cdot)) \le \sum_{k=1}^{r} \rho_K(P(x_{k-1},\cdot), P(x_k,\cdot)). \tag{14.13}$$

Since ρ_K is a minimum over all couplings, the hypotheses of the theorem imply that, for any edge $\{a, b\}$,

$$\rho_K(P(a,\cdot), P(b,\cdot)) \le e^{-\alpha}\ell(a,b). \tag{14.14}$$

Using the bound (14.14) on each of the terms in the sum appearing on the right-hand side of (14.13) shows that

$$\rho_K(P(x,\cdot), P(y,\cdot)) \le e^{-\alpha} \sum_{k=1}^{r} \ell(x_{k-1}, x_k).$$

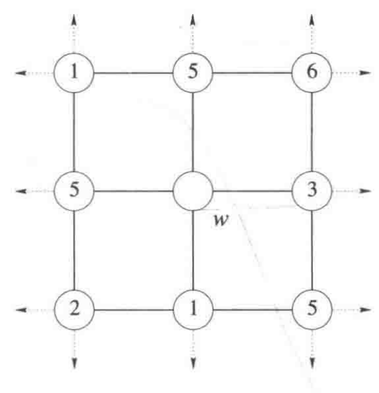
Since the path $(x_0, \dots, x_k)$ was chosen to be of shortest length, the sum on the right-hand side above equals $\rho(x,y)$. This establishes (14.12).

Let η by an optimal coupling of μ and ν, so that

$$\rho_K(\mu,\nu) = \sum_{x,y\in\Omega} \rho(x,y)\eta(x,y). \tag{14.15}$$

By (14.12), we know that for all x, y there exists a coupling $\theta_{x,y}$ of $P(x,\cdot)$ and $P(y,\cdot)$ such that

$$\sum_{u,w\in\Omega} \rho(u,w)\theta_{x,y}(u,w) \le e^{-\alpha}\rho(x,y). \tag{14.16}$$



Colors: {1, 2, 3, 4, 5, 6}

FIGURE 14.1. Updating at vertex w. The colors of the neighbors are not available, as indicated.

Consider the probability distribution $\theta := \sum_{x,y\in\Omega} \eta(x,y)\theta_{x,y}$ on $\Omega \times \Omega$. (This is a coupling of μP with νP.) We have by (14.16) and (14.15) that

$$\begin{aligned}
\sum_{u,w\in\Omega} \rho(u,w)\theta(u,w) &= \sum_{x,y\in\Omega} \sum_{u,w\in\Omega} \rho(u,w)\theta_{x,y}(u,w)\eta(x,y) \\
&\leq e^{-\alpha} \sum_{x,y\in\Omega} \rho(x,y)\eta(x,y) \\
&= e^{-\alpha}\rho_K(\mu,\nu).
\end{aligned}$$

Therefore, the theorem is proved, because $\rho_K(\mu P, \nu P) \leq \sum_{u,w\in\Omega} \rho(u,w)\theta(u,w)$. ■

14.3. Fast Mixing for Colorings

Recall from Section 3.1 that proper q-colorings of a graph $G = (V, E)$ are elements of $x \in \Omega = \{1, 2, \ldots, q\}^V$ such that $x(v) \neq x(w)$ for $\{v, w\} \in E$.

In Section 5.4.1, the mixing time of the Metropolis chain for proper q-colorings was analyzed for sufficiently large q. Here we analyze the mixing time for the Glauber dynamics.

As defined in Section 3.3, Glauber dynamics for proper q-colorings of a graph G with n vertices operate as follows: at each move, a vertex is chosen uniformly at random and the color of this vertex is updated. To update, a color is chosen uniformly at random from the allowable colors, which are those colors not seen among the neighbors of the chosen vertex.

We will use path coupling to bound the mixing time of this chain.

THEOREM 14.8. *Consider the Glauber dynamics chain for random proper q-colorings of a graph with n vertices and maximum degree Δ. If $q > 2\Delta$, then the mixing time satisfies*

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \left(\frac{q-\Delta}{q-2\Delta}\right) n \left(\log n - \log \varepsilon\right) \right\rceil. \tag{14.17}$$

PROOF. The metric here is $\rho(x,y) = \sum_{v \in V} \mathbf{1}\{x(v) \neq y(v)\}$, the number of sites at which x and y differ. Two colorings are neighbors if and only if they differ at a single vertex. Note that this neighboring rule defines a graph different from the graph defined by the transitions of the chain, since the chain moves only among proper colorings.

Recall that $A_v(x)$ is the set of allowable colors at v in configuration x.

Let x and y be two configurations which agree everywhere except at vertex v. We describe how to simultaneously evolve two chains, one started at x and the other started at y, such that each chain viewed alone is a Glauber chain.

First, we pick a vertex w uniformly at random from the vertex set of the graph. (We use a lowercase letter for the random variable w to emphasize that its value is a vertex.) We will update the color of w in both the chain started from x and the chain started from y.

If v is not a neighbor of w, then we can update the two chains with the same color. Each chain is updated with the correct distribution because $A_w(x) = A_w(y)$.

Suppose now one of the neighbors of w is v. We will assume that $|A_w(x)| \leq |A_w(y)|$. If not, run the procedure described below with the roles of x and y reversed.

Generate a random color U from $A_w(y)$ and use this to update y at w. If $U \neq x(v)$, then update the configuration x at w to U. We subdivide the case $U = x(v)$ into subcases based on whether or not $|A_w(x)| = |A_w(y)|$:

case	how to update x at w
$\lvert A_w(x)\rvert = \lvert A_w(y)\rvert$	set $x(w) = y(v)$
$\lvert A_w(x)\rvert < \lvert A_w(y)\rvert$	draw a random color from $A_w(x)$

The reader should check that this updates x at w to a color chosen uniformly from $A_w(x)$. The probability that the two configurations do not update to the same color is $1/|A_w(y)|$, which is bounded above by $1/(q - \Delta)$.

Given two states x and y which are at unit distance (that is, differ in one vertex only), we have constructed a coupling (X_1, Y_1) of $P(x, \cdot)$ and $P(y, \cdot)$. The distance $\rho(X_1, Y_1)$ increases from 1 only in the case where a neighbor of v is updated and the updates are different in the two configurations. Also, the distance decreases when v is selected to be updated. In all other cases the distance stays at 1. Therefore,

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq 1 - \frac{1}{n} + \frac{\deg(v)}{n}\left(\frac{1}{q - \Delta}\right). \tag{14.18}$$

The right-hand side of (14.18) is bounded by

$$1 - \frac{1}{n}\left(1 - \frac{\Delta}{q - \Delta}\right). \tag{14.19}$$

Because $2\Delta < q$, this is not more than 1. Letting $c(q,\Delta) := [1 - \Delta/(q - \Delta)]$,

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq \exp\left(-\frac{c(q,\Delta)}{n}\right).$$

Applying Corollary 14.7 shows that

$$\max_{x \in \Omega} \left\| P^t(x, \cdot) - \pi \right\|_{TV} \leq n \exp\left(-\frac{c(q,\Delta)}{n} t\right)$$

and that

$$t_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \frac{n}{c(q,\Delta)}\left(\log n + \log \varepsilon^{-1}\right) \right\rceil. \tag{14.20}$$

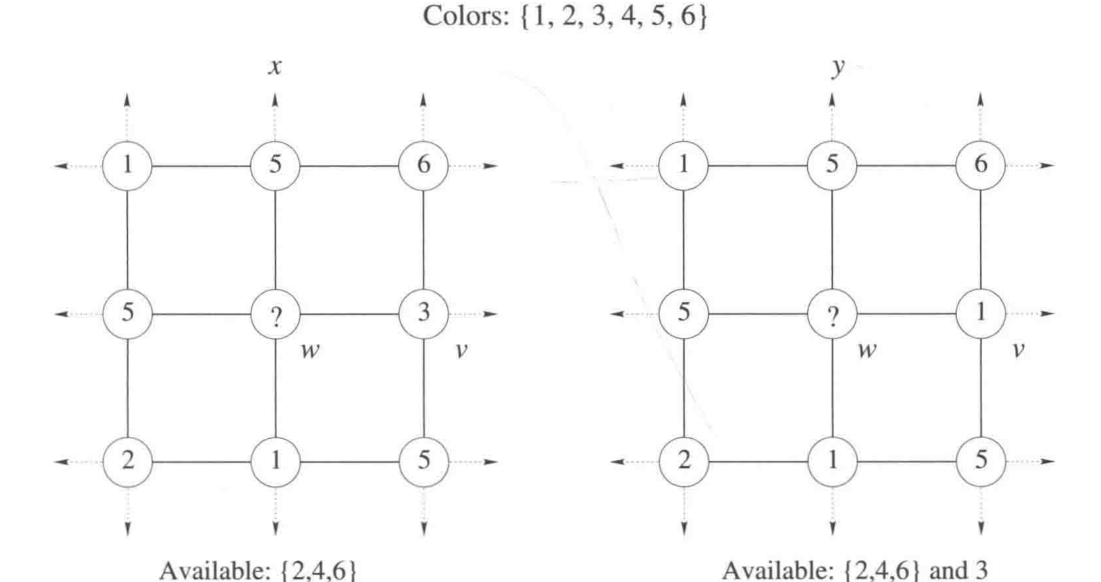


FIGURE 14.2. Jointly updating x and y when they differ only at vertex v and $|A_w(x)| < |A_w(y)|$

(Note that $c(q, \Delta) > 0$ because $q > 2\Delta$.) This establishes (14.17). ∎

Some condition on q and Δ is necessary to achieve the fast rate of convergence (order $n \log n$) established in Theorem 14.8, although the condition $q > 2\Delta$ is not the best known. Example 7.5 shows that if Δ is allowed to grow with n while q remains fixed, then the mixing time can be exponential in n.

Exercise 7.3 shows that for the graph having no edges, in which case the colors at distinct vertices do not interact, the mixing time is at least of order $n \log n$.

14.4. Approximate Counting

14.4.1. Sampling and counting. For sufficiently simple combinatorial sets, it can be easy both to count and to generate a uniform random sample.

EXAMPLE 14.9 (One-dimensional colorings). Recall the definition of proper q-coloring from Section 3.1. On the path with n vertices there are exactly $q(q-1)^{n-1}$ proper colorings: color vertex 1 arbitrarily, and then for each successive vertex $i > 1$, choose a color different from that of vertex $i - 1$. This description of the enumeration is easily modified to a uniform sampling algorithm, as Exercise 14.3 asks the reader to check.

EXAMPLE 14.10 (One-dimensional hardcore model). Now consider the set Ω_n of hardcore configurations on the path with n vertices (recall the definition of the hardcore model in Section 3.3, and see Figure 14.3). Exercise 14.4 asks the reader to check that $|\Omega_n| = f_{n+1}$, where f_n is the n-th Fibonacci number, and Exercise 14.5 asks the reader to check that the following recursive algorithm inductively generates a uniform sample from Ω_n: suppose you are able to generate uniform samples from Ω_k for $k \leq n-1$. With probability f_{n-1}/f_{n+1}, put a 1 at location n, a 0 at location

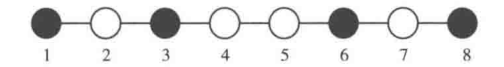


FIGURE 14.3. A configuration of the hardcore model on the 8-vertex path. Filled circles correspond to occupied sites.

$n-1$, and then generate a random element of Ω_{n-2} to fill out the configuration at $\{1, 2, \ldots, n-2\}$. With the remaining probability f_n/f_{n+1}, put a 0 at location n and fill out the positions $\{1, 2, \ldots, n-1\}$ with a random element of Ω_{n-1}.

REMARK 14.11. For more examples of sets enumerated by the Fibonacci numbers, see Stanley (1986, Chapter 1, Exercise 14) and Section 6.6 of Graham, Knuth, and Patashnik (1994). Benjamin and Quinn (2003) use combinatorial interpretations to prove Fibonacci identities (and many other things).

For both models, both sampling and counting become more difficult on more complicated graphs. Fortunately, Markov chains (such as the Glauber dynamics for both these examples) can efficiently sample large combinatorial sets which (unlike the elementary methods described above and in greater generality in Appendix B) do not require enumerating the set or even knowing how many elements are in the set. In Section 14.4.2 we show how Markov chains can be used in an approximate counting algorithm for colorings.

14.4.2. Approximately counting colorings.

Many innovations in the study of mixing times for Markov chains came from researchers motivated by the problem of *counting* combinatorial structures. While determining the exact size of a complicated set may be a "hard" problem, an approximate answer is often possible using Markov chains.

In this section, we show how the number of proper colorings can be estimated using the Markov chain analyzed in the previous section. We adapt the method described in Jerrum and Sinclair (1996) to this setting.

THEOREM 14.12. *Let Ω be the set of all proper q-colorings of the graph G of n vertices and maximal degree Δ. Fix $q > 2\Delta$, and set $c(q, \Delta) = 1 - \Delta/(q - \Delta)$. Given η and ε, there is a random variable W which can be simulated using no more than*

$$\left\lceil \frac{n \log n + n \log(3n/\varepsilon)}{c(q, \Delta)} \right\rceil \left\lceil \frac{27qn^2}{\eta\varepsilon^2} \right\rceil \tag{14.21}$$

uniform random variables and which satisfies

$$\mathbf{P}\{(1-\varepsilon)|\Omega|^{-1} \leq W \leq (1+\varepsilon)|\Omega|^{-1}\} \geq 1 - \eta.$$

REMARK 14.13. This is an example of a ***fully polynomial randomized approximation scheme***, an algorithm for approximating values of the function $n \mapsto |\Omega_n|$ having a run-time that is polynomial in both the *instance size* n and the inverse error tolerated, ε^{-1}.

PROOF. This proof follows closely the argument of Jerrum and Sinclair (1996). Let x_0 be a proper coloring of G. Enumerate the vertices of G as $\{v_1, v_2, \ldots, v_n\}$. Define for $k = 0, 1, \ldots, n$

$$\Omega_k = \{x \in \Omega : x(v_j) = x_0(v_j) \text{ for } j > k\}.$$

Elements of Ω_k have k free vertices, while the $n-k$ vertices $\{v_{k+1},\ldots,v_n\}$ are colored in agreement with x_0.

A random element of Ω_k can be generated using a slight modification to the Glauber dynamics introduced in Section 3.3.1 and analyzed in Section 14.3. The chain evolves as before, but only the colors at vertices $\{v_1,\ldots,v_k\}$ are permitted to be updated. The other vertices are frozen in the configuration specified by x_0. The bound of Theorem 14.8 on $t_{\text{mix}}(\varepsilon)$ still holds, with k replacing n. In addition, (14.20) itself holds, since $k \leq n$. By definition of $t_{\text{mix}}(\varepsilon)$, if

$$t(n,\varepsilon) := \left\lceil \frac{n\log n + n\log(3n/\varepsilon)}{c(q,\Delta)} \right\rceil,$$

then

$$\left\| P^{t(n,\varepsilon)}(x_0,\cdot) - \pi_k \right\|_{\text{TV}} < \frac{\varepsilon}{3n}, \tag{14.22}$$

where π_k is uniform on Ω_k.

The ratio $|\Omega_{k-1}|/|\Omega_k|$ can be estimated as follows: a random element from Ω_k can be generated by running the Markov chain for $t(n,\varepsilon)$ steps. Repeating $a_n := 27qn/\eta\varepsilon^2$ times yields a_n elements of Ω_k. Let $Z_{k,i}$, for $i=1,\ldots,a_n$, be the indicator that the i-th sample is an element of Ω_{k-1}. (Observe that to check if an element x of Ω_k is also an element of Ω_{k-1}, it is enough to determine if $x(v_k) = x_0(v_k)$.) If $x_{k,i}$ is the starting location of the chain used to generate the i-th sample, then

$$|\mathbf{E}Z_{k,i} - \pi_k(\Omega_{k-1})| = |P^{t(n,\varepsilon)}(x_{k,i},\Omega_{k-1}) - \pi_k(\Omega_{k-1})| \leq d(t(n,\varepsilon)) \leq \frac{\varepsilon}{3n}.$$

Therefore, if $W_k := a_n^{-1}\sum_{i=1}^{a_n} Z_{k,i}$ is the fraction of these samples which fall in Ω_{k-1}, then

$$\left| \mathbf{E}W_k - \frac{|\Omega_{k-1}|}{|\Omega_k|} \right| \leq \frac{1}{a_n}\sum_{i=1}^{a_n} |\mathbf{E}Z_{k,i} - \pi_k(\Omega_{k-1})| \leq \frac{\varepsilon}{3n}. \tag{14.23}$$

Because $Z_{k,i}$ is a Bernoulli($\mathbf{E}Z_{k,i}$) random variable and the $Z_{k,i}$'s are independent,

$$\mathrm{Var}(W_k) = \frac{1}{a_n^2}\sum_{i=1}^{a_n} \mathbf{E}Z_{k,i}[1-\mathbf{E}Z_{k,i}] \leq \frac{1}{a_n^2}\sum_{i=1}^{a_n}\mathbf{E}Z_{k,i} \leq \frac{\mathbf{E}(W_k)}{a_n}.$$

Consequently,

$$\frac{\mathrm{Var}(W_k)}{\mathbf{E}^2(W_k)} \leq \frac{1}{a_n\mathbf{E}(W_k)}. \tag{14.24}$$

Since $|\Omega_{k-1}|/|\Omega_k| \geq q^{-1}$, we have from (14.23) that, for n large enough,

$$\mathbf{E}(W_k) \geq \frac{1}{q} - \frac{\varepsilon}{3n} \geq \frac{1}{2q}.$$

Using the above in (14.24) shows that

$$\frac{\mathrm{Var}(W_k)}{\mathbf{E}^2(W_k)} \leq \frac{2q}{a_n} = \frac{2\eta\varepsilon^2}{27n}. \tag{14.25}$$

Recall the inequality $|\prod z_i - \prod w_i| \le \sum |z_i - w_i|$, valid for $|z_i| \le 1$ and $|w_i| \le 1$. Letting $W = W_1 \cdots W_n$, since the $\{W_k\}$ are independent, we have

$$\begin{aligned}\left|\mathbf{E}(W) - \frac{1}{|\Omega|}\right| &= \left|\prod_{i=1}^{n} \mathbf{E}W_i - \prod_{i=1}^{n} \frac{|\Omega_{i-1}|}{|\Omega_i|}\right| \\ &\le \sum_{i=1}^{n} \left|\mathbf{E}W_i - \frac{|\Omega_{i-1}|}{|\Omega_i|}\right| \le n \cdot \frac{\varepsilon}{3n} = \frac{\varepsilon}{3}.\end{aligned}$$

Therefore,

$$\mathbf{E}(W) = \frac{1}{|\Omega|} + \tilde{\varepsilon}, \quad \text{where } |\tilde{\varepsilon}| \le \varepsilon/3. \tag{14.26}$$

Also,

$$\mathbf{E}\left(\frac{W}{\mathbf{E}W}\right)^2 = \mathbf{E}\prod_{i=1}^{n}\left(\frac{W_i}{\mathbf{E}W_i}\right)^2 = \prod_{i=1}^{n} \frac{\mathbf{E}W_i^2}{(\mathbf{E}W_i)^2}.$$

Subtracting 1 from both sides shows that

$$\frac{\mathrm{Var}(W)}{\mathbf{E}^2(W)} = \prod_{k=1}^{n}\left[1 + \frac{\mathrm{Var}\, W_k}{\mathbf{E}^2(W_k)}\right] - 1.$$

This identity, together with (14.25), shows that

$$\frac{\mathrm{Var}(W)}{\mathbf{E}^2(W)} \le \prod_{k=1}^{n}\left[1 + \frac{2\eta\varepsilon^2}{27n}\right] - 1 \le \frac{\eta\varepsilon^2}{9}.$$

By Chebyshev's inequality,

$$\mathbf{P}\left\{\left|\frac{W}{\mathbf{E}(W)} - 1\right| \ge \varepsilon/3\right\} \le \eta.$$

Combining with (14.26),

$$\mathbf{P}\left\{\left|\frac{W}{|\Omega|^{-1}} - 1\right| \ge \varepsilon\right\} \le \eta.$$

For each of the n variables W_k, $k = 1, \ldots, n$, we need to simulate each of a_n chains for $t(n, \varepsilon)$ steps. This shows that a total of (14.21) steps are needed. ■

Exercises

EXERCISE 14.1. Let M be an arbitrary set, and, for $a, b \in M$, define

$$\rho(a, b) = \begin{cases} 0 & \text{if } a = b, \\ 1 & \text{if } a \ne b. \end{cases} \tag{14.27}$$

Check that M is a metric space under the distance ρ and the corresponding transportation metric is the total variation distance.

EXERCISE 14.2. A real-valued function f on a metric space (Ω, ρ) is called ***Lipschitz*** if there is a constant c so that for all $x, y \in \Omega$,

$$|f(x) - f(y)| \le c\rho(x, y), \tag{14.28}$$

where ρ is the distance on Ω. We denote the best constant c in (14.28) by $\mathrm{Lip}(f)$:

$$\mathrm{Lip}(f) := \max_{\substack{x, y \in \Omega \\ x \ne y}} \frac{|f(x) - f(y)|}{\rho(x, y)}.$$

For a probability μ on Ω, the integral $\int f d\mu$ denotes the sum $\sum_{x\in\Omega} f(x)\mu(x)$. Define

$$\tilde{\rho}_k(\mu,\nu) = \sup_{f\,:\,\mathrm{Lip}(f)\leq 1} \left| \int f d\mu - \int f d\nu \right|.$$

Show that $\tilde{\rho}_K \leq \rho_K$. (In fact, $\tilde{\rho}_K = \rho_K$; see Notes.)

EXERCISE 14.3. Let $H(1)$ be a uniform sample from $[k]$. Given that $H(i)$ has been assigned for $i = 1, \ldots, j-1$, choose $H(j)$ uniformly from $[k] \setminus \{H(j-1)\}$. Repeat for $j = 2, \ldots, n$. Show that H is a uniform sample from the set of colorings of the n-vertex path.

EXERCISE 14.4. Recall that the ***Fibonacci numbers*** are defined by $f_0 := f_1 := 1$ and $f_n := f_{n-1} + f_{n-2}$ for $n \geq 1$. Show that the number of configurations in the one-dimensional hardcore model with n sites is f_{n+1}.

EXERCISE 14.5. Show that the algorithm described in Example 14.10 generates a uniform sample from Ω_n.

EXERCISE 14.6. Describe a simple exact sampling mechanism, in the style of Exercises 14.3 and 14.5, for the Ising model on the n-vertex path.

Notes

The transportation metric was introduced in Kantorovich (1942). It has been rediscovered many times and is also known as the Wasserstein metric, thanks to a reintroduction in Vasershtein (1969). For some history of this metric, see Vershik (2004). See also Villani (2003).

The name "transportation metric" comes from the following problem: suppose a unit of materiel is spread over n locations $\{1, 2, \ldots, n\}$ according to the distribution μ, so that proportion $\mu(i)$ is at location i. You wish to re-allocate the materiel according to another distribution ν, and the per unit cost of moving from location i to location j is $\rho(i,j)$. For each i and j, what proportion $p(i,j)$ of mass at location i should be moved to location j so that $\sum_{i=1}^{n} \mu(i)p(i,j)$, the total amount moved to location j, equals $\nu(j)$ and so that the total cost is minimized? The total cost when using p equals

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \rho(i,j)\mu(i)p(i,j).$$

Since $q(i,j) = \mu(i)p(i,j)$ is a coupling of μ and ν, the problem is equivalent to finding the coupling q which minimizes

$$\sum_{1\leq i,j\leq n} \rho(i,j)q(i,j).$$

The problem of mixing for chains with stationary distribution uniform over proper q-colorings was first analyzed by Jerrum (1995), whose bound we present as Theorem 14.8, and independently by Salas and Sokal (1997). Vigoda (2000) showed that if the number of colors q is larger than $(11/6)\Delta$, then the mixing times for the Glauber dynamics for random colorings is $O(n^2 \log n)$. Dyer, Greenhill, and Molloy (2002) show that the mixing time is $O(n \log n)$ provided $q \geq (2 - 10^{-12})\Delta$. A key open question is whether $q > \Delta + C$ suffices for rapid mixing. Frieze and Vigoda (2007) wrote a survey on using Markov chains to sample from colorings.

The inequality in Exercise 14.2 is actually an equality, as was shown in Kantorovich and Rubinstein (1958). In fact, the theorem is valid more generally on separable metric spaces; the proof uses a form of duality. See Dudley (2002, Theorem 11.8.2).

The relation between sampling and approximate counting first appeared in Jerrum, Valiant, and Vazirani (1986). Jerrum, Sinclair, and Vigoda (2004) approximately count perfect matchings in bipartite graphs. For more on approximate counting, see Sinclair (1993).

CHAPTER 15

The Ising Model

The Ising model on a graph with vertex set V at inverse temperature β was introduced in Section 3.3.5. It is the probability distribution on $\Omega = \{-1,1\}^V$ defined by

$$\pi(\sigma) = Z(\beta)^{-1} \exp\left(\beta \sum_{\substack{v,w \in V \\ v \sim w}} \sigma(v)\sigma(w)\right).$$

Here we study in detail the Glauber dynamics for this distribution. As discussed in Section 3.3.5, the transition matrix for this chain is given by

$$P(\sigma,\sigma') = \frac{1}{n}\sum_{v\in V} \frac{e^{\beta\,\sigma'(v)\,S(\sigma,v)}}{e^{\beta\,\sigma'(v)\,S(\sigma,v)} + e^{-\beta\,\sigma'(v)\,S(\sigma,v)}} \cdot \mathbf{1}_{\{\sigma'(w)=\sigma(w) \text{ for all } w\neq v\}},$$

where $S(\sigma,v) = \sum_{w\,:\,w\sim v} \sigma(w)$.

This chain evolves by selecting a vertex v at random and updating the spin at v according to the distribution π conditioned to agree with the spins at all vertices not equal to v. If the current configuration is σ and vertex v is selected, then the chance the spin at v is updated to $+1$ is equal to

$$p(\sigma,v) := \frac{e^{\beta S(\sigma,v)}}{e^{\beta S(\sigma,v)} + e^{-\beta S(\sigma,v)}} = \frac{1+\tanh(\beta S(\sigma,v))}{2}.$$

We will be particularly interested in how the mixing time varies with β. Generically, for small values of β, the chain will mix in a short amount of time, while for large values of β, the chain will converge slowly. Understanding this phase transition between slow and fast mixing has been a topic of great interest and activity over the past twenty years; here we only scratch the surface.

15.1. Fast Mixing at High Temperature

In this section we use the path coupling technique of Chapter 14 to show that on any graph of bounded degree, for small values of β, the Glauber dynamics for the Ising model is fast mixing.

THEOREM 15.1. *Consider the Glauber dynamics for the Ising model on a graph with n vertices and maximal degree Δ.*

(i) *Let $c(\beta) := 1 - \Delta\tanh(\beta)$. If $\Delta\cdot\tanh(\beta) < 1$, then*

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{n(\log n + \log(1/\varepsilon))}{c(\beta)} \right\rceil. \tag{15.1}$$

In particular, (15.1) *holds whenever $\beta < \Delta^{-1}$.*

(ii) *Suppose every vertex of the graph has even degree. Let*

$$c_e(\beta) := 1 - (\Delta/2)\tanh(2\beta).$$

If $(\Delta/2)\cdot\tanh(2\beta) < 1$, *then*

$$t_{\mathrm{mix}}(\varepsilon) \le \left\lceil \frac{n(\log n + \log(1/\varepsilon))}{c_e(\beta)} \right\rceil. \tag{15.2}$$

LEMMA 15.2. *The function* $\varphi(x) := \tanh(\beta(x+1)) - \tanh(\beta(x-1))$ *is even and decreasing on* $[0,\infty)$, *whence*

$$\sup_{x\in\mathbb{R}} \varphi(x) = \varphi(0) = 2\tanh(\beta) \tag{15.3}$$

and

$$\sup_{k \text{ odd integer}} \varphi(k) = \varphi(1) = \tanh(2\beta). \tag{15.4}$$

PROOF. Let $\psi(x) := \tanh(\beta x)$; observe that $\psi'(x) = \beta/\cosh^2(\beta x)$. The function ψ' is strictly positive and decreasing on $[0,\infty)$ and is even. Therefore, for $x > 0$,

$$\varphi'(x) = \psi'(x+1) - \psi'(x-1) < 0,$$

as is seen by considering separately the case where $x - 1 > 0$ and the case where $x - 1 \le 0$. Because tanh is an odd function,

$$\varphi(-x) = \psi(-x+1) - \psi(-x-1) = -\psi(x-1) + \psi(x+1) = \varphi(x),$$

so φ is even. ∎

PROOF OF THEOREM 15.1. Define the distance ρ on Ω by

$$\rho(\sigma,\tau) = \frac{1}{2}\sum_{u\in V} |\sigma(u) - \tau(u)|.$$

The distance ρ is a path metric as defined in Section 14.2.

Let σ and τ be two configurations with $\rho(\sigma,\tau) = 1$. The spins of σ and τ agree everywhere except at a single vertex v. Assume that $\sigma(v) = -1$ and $\tau(v) = +1$.

Define $\mathcal{N}(v) := \{u : u \sim v\}$ to be the set of neighboring vertices to v.

We describe now a coupling (X, Y) of one step of the chain started in configuration σ with one step of the chain started in configuration τ.

Pick a vertex w uniformly at random from V. If $w \notin \mathcal{N}(v)$, then the neighbors of w agree in both σ and τ. As the probability of updating the spin at w to $+1$, given in (3.10), depends only on the spins at the neighbors of w, it is the same for the chain started in σ as for the chain started in τ. Thus we can update both chains together.

If $w \in \mathcal{N}(v)$, the probabilities of updating to $+1$ at w are no longer the same for the two chains, so we cannot *always* update together. We do, however, use a single random variable as the common source of noise to update both chains, so the two chains agree as often as is possible. In particular, let U be a uniform random variable on $[0,1]$ and set

$$X(w) = \begin{cases} +1 & \text{if } U \le p(\sigma, w), \\ -1 & \text{if } U > p(\sigma, w) \end{cases} \quad \text{and} \quad Y(w) = \begin{cases} +1 & \text{if } U \le p(\tau, w), \\ -1 & \text{if } U > p(\tau, w). \end{cases}$$

Set $X(u) = \sigma(u)$ and $Y(u) = \tau(u)$ for $u \ne w$.

If $w = v$, then $\rho(X,Y) = 0$. If $w \notin \mathcal{N}(v) \cup \{v\}$, then $\rho(X,Y) = 1$. If $w \in \mathcal{N}(v)$ and $p(\sigma, w) < U \leq p(\tau, w)$, then $\rho(X,Y) = 2$. Thus,

$$\mathbf{E}_{\sigma,\tau}(\rho(X,Y)) \leq 1 - \frac{1}{n} + \frac{1}{n} \sum_{w \in \mathcal{N}(v)} [p(\tau,w) - p(\sigma,w)]. \tag{15.5}$$

Noting that $S(w,\tau) = S(w,\sigma) + 2 = S + 2$, we obtain

$$\begin{aligned} p(\tau,w) - p(\sigma,w) &= \frac{e^{\beta(S+2)}}{e^{\beta(S+2)} + e^{-\beta(S+2)}} - \frac{e^{\beta S}}{e^{\beta S} + e^{-\beta S}} \\ &= \frac{1}{2}\left[\tanh(\beta(S+2)) - \tanh(\beta S)\right]. \end{aligned} \tag{15.6}$$

Letting $\tilde{S} = S + 1$ in (15.6) and then applying (15.3) shows that

$$p(\tau,w) - p(\sigma,w) = \frac{1}{2}\left[\tanh(\beta(\tilde{S}+1)) - \tanh(\beta(\tilde{S}-1))\right] \leq \tanh(\beta). \tag{15.7}$$

Using the above bound in inequality (15.5) shows that

$$\mathbf{E}_{\sigma,\tau}\left(\rho(X,Y)\right) \leq 1 - \frac{[1 - \Delta\tanh(\beta)]}{n} \leq \exp\left(-\frac{1-\Delta\tanh(\beta)}{n}\right) = e^{-c(\beta)/n}.$$

If $\Delta \tanh(\beta) < 1$, then $c(\beta) > 0$. Observe that $\operatorname{diam}(\Omega) = n$. Applying Corollary 14.7 with $\alpha = c(\beta)/n$ establishes (15.1).

Since $\tanh(x) \leq x$, if $\beta < \Delta^{-1}$, then $\Delta\tanh(\beta) < 1$.

Proof of (ii). Note that if every vertex in the graph has even degree, then $\tilde{S} = S + 1$ takes on only odd values. Applying (15.4) shows that

$$p(\tau,w) - p(\sigma,w) = \frac{1}{2}\left[\tanh(\beta(\tilde{S}+1)) - \tanh(\beta(\tilde{S}-1))\right] \leq \frac{\tanh(2\beta)}{2}.$$

Using the above bound in inequality (15.5) shows that

$$\mathbf{E}_{\sigma,\tau}\left(\rho(X,Y)\right) \leq 1 - \frac{1 - (\Delta/2)\tanh(2\beta)}{n} \leq e^{-c_e(\beta)/n}.$$

If $(\Delta/2)\tanh(2\beta) < 1$, then we can apply Corollary 14.7 to obtain (15.2). ∎

15.2. The Complete Graph

Let G be the complete graph on n vertices, the graph which includes all $\binom{n}{2}$ possible edges. Since the interaction term $\sigma(v)\sum_{w\,:\,w\sim v}\sigma(w)$ is of order n, we take $\beta = \alpha/n$ so that the total contribution of a single site to $\beta\sum\sigma(v)\sigma(w)$ is $O(1)$.

THEOREM 15.3. *Let G be the complete graph on n vertices, and consider Glauber dynamics for the Ising model on G with $\beta = \alpha/n$.*

(i) *If $\alpha < 1$, then*

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{n(\log n + \log(1/\varepsilon))}{1-\alpha}. \tag{15.8}$$

(ii) *If $\alpha > 1$, then there is a positive function $r(\alpha)$ so that $t_{\mathrm{mix}} \geq O\left(\exp\left[r(\alpha)n\right]\right)$.*

PROOF. *Proof of* (i). Note that $\Delta\tanh(\beta) = (n-1)\tanh(\alpha/n) \leq \alpha$. Thus if $\alpha < 1$, then Theorem 15.1(i) establishes (15.8).

Proof of (ii). Define $A_k := \{\sigma : |\{v : \sigma(v) = 1\}| = k\}$. By counting, $\pi(A_k) = a_k/Z(\alpha)$, where

$$a_k := \binom{n}{k}\exp\left\{\frac{\alpha}{n}\left[\binom{k}{2} + \binom{n-k}{2} - k(n-k)\right]\right\}.$$

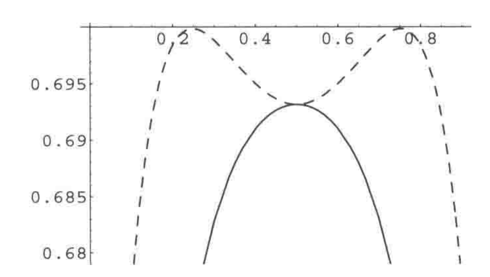


FIGURE 15.1. The function φ_α defined in (15.9). The dashed graph corresponds to $\alpha = 1.1$, the solid line to $\alpha = 0.9$.

Taking logarithms and applying Stirling's formula shows that

$$\log(a_{\lfloor cn \rfloor}) = n\varphi_\alpha(c)[1 + o(1)],$$

where

$$\varphi_\alpha(c) := -c\log(c) - (1-c)\log(1-c) + \alpha\left[\frac{(1-2c)^2}{2}\right]. \tag{15.9}$$

Taking derivatives shows that

$$\varphi'_\alpha(1/2) = 0,$$
$$\varphi''_\alpha(1/2) = -4(1-\alpha).$$

Hence $c = 1/2$ is a critical point of φ_α, and in particular it is a local maximum or minimum depending on the value of α. See Figure 15.1 for the graph of φ_α for $\alpha = 0.9$ and $\alpha = 1.1$. Take $\alpha > 1$, in which case φ_α has a local minimum at $1/2$. Define

$$S = \left\{\sigma : \sum_{u \in V} \sigma(u) < 0\right\}.$$

By symmetry, $\pi(S) \leq 1/2$. Observe that the only way to get from S to S^c is through $A_{\lfloor n/2 \rfloor}$, since we are only allowed to change one spin at a time. Thus

$$Q(S, S^c) \leq \frac{\lceil (n/2) \rceil}{n}\pi(A_{\lceil n/2 \rceil}) \quad \text{and} \quad \pi(S) = \sum_{j < \lceil n/2 \rceil} \pi(A_j).$$

Let c_1 be the value of c maximizing φ_α over $[0, 1/2]$. Since $1/2$ is a strict local minimum, $c_1 < 1/2$. Therefore,

$$\Phi(S) \leq \frac{\exp\{\varphi_\alpha(1/2)n[1+o(1)]\}}{Z(\alpha)\pi(A_{\lfloor c_1 n \rfloor})} = \frac{\exp\{\varphi_\alpha(1/2)n[1+o(1)]\}}{\exp\{\varphi_\alpha(c_1)n[1+o(1)]\}}.$$

Since $\varphi_\alpha(c_1) > \varphi_\alpha(1/2)$, there is an $r(\alpha) > 0$ and constant $b > 0$ so that $\Phi_\star \leq be^{-nr(\alpha)}$. The conclusion follows from Theorem 7.3. ■

15.3. The Cycle

THEOREM 15.4. *Let $c_O(\beta) := 1 - \tanh(2\beta)$. The Glauber dynamics for the Ising model on the n-cycle satisfies, for any $\beta > 0$ and fixed $\varepsilon > 0$,*

$$\frac{1 + o(1)}{2c_O(\beta)} \leq \frac{t_{\text{mix}}(\varepsilon)}{n\log n} \leq \frac{1+o(1)}{c_O(\beta)}. \tag{15.10}$$

PROOF. *Upper bound.* Note that $\Delta = 2$, whence $(\Delta/2)\tanh(2\beta) = \tanh(2\beta) < 1$. Since the degree of every vertex in the cycle is two, Theorem 15.1(ii) shows that

$$t_{\mathrm{mix}}(\varepsilon) \le \frac{n(\log n + \log(1/\varepsilon))}{c_{\mathrm{O}}(\beta)}$$

for all β.

Lower bound. We will use Wilson's method (Theorem 13.5).

Claim: The function $\Phi : \Omega \to \mathbb{R}$ defined by $\Phi(\sigma) := \sum_{i=1}^{n} \sigma(i)$ is an eigenfunction with eigenvalue

$$\lambda = 1 - \frac{1 - \tanh(2\beta)}{n}. \tag{15.11}$$

Proof of Claim: We first consider the action of P on $\varphi_i : \Omega \to \mathbb{R}$ defined by $\varphi_i(\sigma) := \sigma_i$. Recall that if vertex i is selected for updating, a positive spin is placed at i with probability

$$\frac{1 + \tanh\left[\beta(\sigma(i-1) + \sigma(i+1))\right]}{2}.$$

(Cf. (3.10); here $S(\sigma, i) = \sum_{j\,:\,j\sim i} \sigma(j) = \sigma(i-1) + \sigma(i+1)$.) Therefore,

$$\begin{aligned}(P\varphi_i)(\sigma) &= (+1)\left(\frac{1 + \tanh[\beta(\sigma(i-1) + \sigma(i+1))]}{2n}\right) \\ &\quad + (-1)\left(\frac{1 - \tanh[\beta(\sigma(i-1) + \sigma(i+1))]}{2n}\right) + \left(1 - \frac{1}{n}\right)\sigma(i) \\ &= \frac{\tanh[\beta(\sigma(i-1) + \sigma(i+1))]}{n} + \left(1 - \frac{1}{n}\right)\sigma(i).\end{aligned}$$

The variable $[\sigma(i-1) + \sigma(i+1)]$ takes values in $\{-2, 0, 2\}$; since the function tanh is odd, it is linear on $\{-2, 0, 2\}$ and in particular, for $x \in \{-2, 0, 2\}$,

$$\tanh(\beta x) = \frac{\tanh(2\beta)}{2} x.$$

We conclude that

$$(P\varphi_i)(\sigma) = \frac{\tanh(2\beta)}{2n}\left(\sigma(i-1) + \sigma(i+1)\right) + \left(1 - \frac{1}{n}\right)\sigma(i).$$

Summing over i,

$$(P\Phi)(\sigma) = \frac{\tanh(2\beta)}{n}\Phi(\sigma) + \left(1 - \frac{1}{n}\right)\Phi(\sigma) = \left(1 - \frac{1 - \tanh(2\beta)}{n}\right)\Phi(\sigma),$$

proving that Φ is an eigenfunction with eigenvalue λ defined in (15.11).

Note that if $\tilde{\sigma}$ is the state obtained after updating σ according to the Glauber dynamics, then $|\Phi(\tilde{\sigma}) - \Phi(\sigma)| \le 2$. Therefore, taking x to be the configuration of all ones, (13.3) yields

$$\begin{aligned}t_{\mathrm{mix}}(\varepsilon) &\ge [1 + o(1)]\left[\frac{n}{2c_{\mathrm{O}}(\beta)}\left(\log\left(\frac{\frac{c_{\mathrm{O}}(\beta)}{n}n^2}{4}\right) + \log\left(\frac{1}{2\varepsilon}\right)\right)\right] \\ &= \frac{[1 + o(1)]n\log n}{2c_{\mathrm{O}}(\beta)}.\end{aligned}$$

■

15.4. The Tree

Our applications of path coupling have heretofore used path metrics with unit edge lengths. Let $\theta := \tanh(\beta)$. The coupling of Glauber dynamics for the Ising model that was used in Theorem 15.1 contracts the Hamming distance, provided $\theta\Delta < 1$. Therefore, the Glauber dynamics for the Ising model on a b-ary tree mixes in $O(n \log n)$ steps, provided $\theta < 1/(b+1)$. We now improve this, showing that the same coupling contracts a weighted path metric whenever $\theta < 1/(2\sqrt{b})$. While this result is not the best possible (see the Notes), it does illustrate the utility of allowing for variable edge lengths in the path metric.

Let T be a finite, rooted b-ary tree of depth k. Fix $0 < \alpha < 1$. We define a graph with vertex set $\{-1,1\}^T$ by placing an edge between configurations σ and τ if they agree everywhere except at a single vertex v. The length of this edge is defined to be $\alpha^{|v|-k}$, where $|v|$ denotes the depth of vertex v. The shortest path between arbitrary configurations σ and τ has length

$$\rho(\sigma,\tau) = \sum_{v\in T} \alpha^{|v|-k} \mathbf{1}_{\{\sigma(v)\neq\tau(v)\}}. \tag{15.12}$$

THEOREM 15.5. *Let $\theta := \tanh(\beta)$. Consider the Glauber dynamics for the Ising model on T, the finite rooted b-ary tree of depth k. If $\alpha = 1/\sqrt{b}$, then for any pair of neighboring configurations σ and τ, there is a coupling (X_1, Y_1) of the Glauber dynamics started from σ and τ such that the metric ρ defined in* (15.12) *contracts when $\theta < 1/(2\sqrt{b})$: there exists a constant $0 < c_\theta < 1$ such that*

$$\mathbf{E}_{\sigma,\tau}[\rho(X_1,Y_1)] \le \left(1 - \frac{c_\theta}{n}\right)\rho(\sigma,\tau).$$

Therefore, if $\theta < 1/(2\sqrt{b})$, then

$$t_{\text{mix}}(\varepsilon) \le \frac{n}{c_\theta}\left[\frac{3}{2}\log n + \log(1/\varepsilon)\right].$$

PROOF. Suppose that σ and τ are configurations which agree everywhere except v, where $-1 = \sigma(v) = -\tau(v)$. Therefore, $\rho(\sigma,\tau) = \alpha^{|v|-k}$. Let (X_1, Y_1) be one step of the coupling used in Theorem 15.1.

We say the coupling ***fails*** if a neighbor w of v is selected and the coupling does not update the spin at w identically in both σ and τ. Given a neighbor of v is selected for updating, the coupling fails with probability

$$p(\tau,w) - p(\sigma,w) \le \theta.$$

(See (15.7).)

If a child w of v is selected for updating and the coupling fails, then the distance increases by

$$\rho(X_1,Y_1) - \rho(\sigma,\tau) = \alpha^{|v|-k+1} = \alpha\rho(\sigma,\tau).$$

If the parent of w is selected for updating and the coupling fails, then the distance increases by

$$\rho(X_1,Y_1) - \rho(\sigma,\tau) = \alpha^{|v|-k-1} = \alpha^{-1}\rho(\sigma,\tau). \tag{15.13}$$

Therefore,

$$\frac{\mathbf{E}_{\sigma,\tau}[\rho(X_1,Y_1)]}{\rho(\sigma,\tau)} \le 1 - \frac{1}{n} + \frac{(\alpha^{-1} + b\alpha)\theta}{n}.$$

The function $\alpha \mapsto \alpha^{-1} + b\alpha$ is minimized over $[0,1]$ at $\alpha = 1/\sqrt{b}$, where it has value $2\sqrt{b}$. Thus, the right-hand side of (15.13), for this choice of α, equals

$$1 - \frac{1 - 2\theta\sqrt{b}}{n}.$$

For $\theta < 1/[2\sqrt{b}]$ we obtain a contraction.

The diameter of the tree in the metric ρ is not more than $\alpha^{-k} n = b^{k/2} n$. Since $b^k < n$, the diameter is at most $n^{3/2}$. Applying Corollary 14.7 completes the proof. ■

We now show that at any temperature, the mixing time on a finite b-ary tree is at most polynomial in the volume of the tree.

THEOREM 15.6 (Kenyon, Mossel, and Peres (2001)). *The Glauber dynamics for the Ising model on the finite, rooted, b-ary tree of depth k satisfies*

$$t_{\text{rel}} \le n_k^{c_T(\beta,b)},$$

where $c_T(\beta, b) := 2\beta(3b+1)/\log b + 1$ *and* n_k *is the number of vertices in the tree.*

To prove Theorem 15.6, we first need a proposition showing the effect on the dynamics of removing an edge of the underlying graph. The following applies more generally than for trees.

PROPOSITION 15.7. *Let* $G = (V, E)$ *have maximal degree* Δ, *where* $|V| = n$, *and let* $\tilde{G} = (V, \tilde{E})$, *where* $\tilde{E} \subset E$. *Let* $r = |E \setminus \tilde{E}|$. *If* γ *is the spectral gap for Glauber dynamics for the Ising model on* G *and* $\tilde{\gamma}$ *is the spectral gap for the dynamics on* $\tilde{G}$, *then*

$$\frac{1}{\gamma} \le \frac{e^{2\beta(\Delta+2r)}}{\tilde{\gamma}}$$

PROOF. We have

$$\begin{aligned}\pi(\sigma) &= \frac{e^{\beta\sum_{(v,w)\in\tilde{E}}\sigma(v)\sigma(w) + \beta\sum_{(v,w)\in E\setminus\tilde{E}}\sigma(v)\sigma(w)}}{\sum_\tau e^{\beta\sum_{(v,w)\in\tilde{E}}\tau(v)\tau(w) + \beta\sum_{(v,w)\in E\setminus\tilde{E}}\tau(v)\tau(w)}}\\ &\ge \frac{e^{-\beta r}}{e^{\beta r}} \frac{e^{\beta\sum_{(v,w)\in\tilde{E}}\sigma(v)\sigma(w)}}{\sum_\tau e^{\beta\sum_{(v,w)\in\tilde{E}}\tau(v)\tau(w)}}\\ &= e^{-2\beta r}\tilde{\pi}(\sigma).\end{aligned}$$

Therefore,

$$\tilde{\pi}(\sigma) \le e^{2\beta r}\pi(\sigma). \tag{15.14}$$

Since the transition matrix is given by (3.11), we have

$$P(\sigma,\tau) \ge \frac{1}{n}\frac{1}{1+e^{2\beta\Delta}}\mathbf{1}\{P(\sigma,\tau) > 0\}$$

and also

$$\tilde{P}(\sigma,\tau) \le \frac{1}{n}\frac{e^{2\beta\Delta}}{1+e^{2\beta\Delta}}\mathbf{1}\{P(\sigma,\tau) > 0\}.$$

Combining these two inequalities shows that $\tilde{P}(\sigma,\tau) \le e^{2\beta\Delta}P(\sigma,\tau)$, whence by (15.14) we have

$$\tilde{\pi}(\sigma)\tilde{P}(\sigma,\tau) \le e^{2\beta(\Delta+r)}\pi(\sigma)P(\sigma,\tau),$$

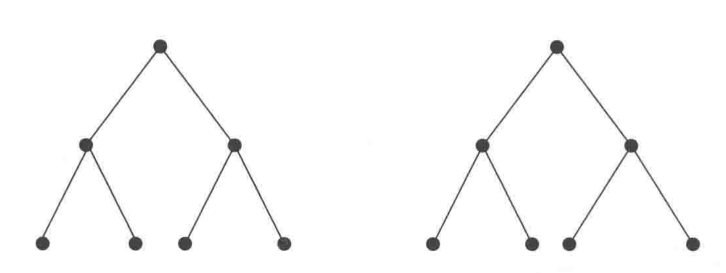


FIGURE 15.2. The tree $\tilde{T}_{2,3}$.

and by (13.11), $\tilde{\mathcal{E}}(f) \leq e^{2\beta(\Delta+r)}\mathcal{E}(f)$. Since $\pi(\sigma) \leq e^{2\beta r}\tilde{\pi}(\sigma)$ (as seen by reversing the roles of π and $\tilde{\pi}$ in the proof of (15.14)), by Lemma 13.22 we have that

$$\tilde{\gamma} \leq e^{2\beta(\Delta+2r)}\gamma.$$

■

PROOF OF THEOREM 15.6. Let $\tilde{T}_{b,k}$ be the graph obtained by removing all edges incident to the root. (See Figure 15.2.)

By Proposition 15.7,

$$\frac{t_{\mathrm{rel}}(T_{k+1})}{n_{k+1}} \leq e^{2\beta(3b+1)}\frac{t_{\mathrm{rel}}(\tilde{T}_{b,k+1})}{n_{k+1}}.$$

Applying Lemma 12.13 shows that

$$\frac{t_{\mathrm{rel}}(\tilde{T}_{b,k+1})}{n_{k+1}} = \max\left\{1, \frac{t_{\mathrm{rel}}(T_k)}{n_k}\right\}.$$

Therefore, if $t_k := t_{\mathrm{rel}}(T_k)/n_k$, then $t_{k+1} \leq e^{2\beta(3b+1)}\max\{t_k, 1\}$. We conclude that, since $n_k \geq b^k$,

$$t_{\mathrm{rel}}(T_k) \leq e^{2\beta(3b+1)k}n_k = (b^k)^{2\beta(3b+1)/\log b}n_k \leq n_k^{2\beta(3b+1)/\log b+1}.$$

■

REMARK 15.8. The proof of Theorem 15.6 shows the utility of studying product systems. Even though the dynamics on the tree does not have independent components, it can be compared to the dynamics on disjoint components, which has product form.

15.5. Block Dynamics

Let $V_i \subset V$ for $i = 1, \ldots, b$ be subsets of vertices, which we will refer to as ***blocks***. The ***block dynamics*** for the Ising model is the Markov chain defined as follows: a block V_i is picked uniformly at random among the b blocks, and the configuration σ is updated according to the measure π conditioned to agree with σ everywhere outside of V_i. More precisely, for $W \subset V$ let

$$\Omega_{\sigma,W} := \{\tau \in \Omega : \tau(v) = \sigma(v) \text{ for all } v \notin W\}$$

be the set of configurations agreeing with σ outside of W, and define the transition matrix

$$P_W(\sigma, \tau) := \pi(\tau \mid \Omega_{\sigma,W}) = \frac{\pi(\tau)\mathbf{1}_{\{\tau \in \Omega_{\sigma,W}\}}}{\pi(\Omega_{\sigma,W})}.$$

The block dynamics has transition matrix $\tilde{P} := b^{-1} \sum_{i=1}^{n} P_{V_i}$.

THEOREM 15.9. *Consider the block dynamics for the Ising model, with blocks* $\{V_i\}_{i=1}^{b}$. *Let* $M := \max_{1 \le i \le b} |V_i|$, *and let* $M^\star := \max_{v \in V} |\{i : v \in V_i\}|$. *Assume that* $\bigcup_{i=1}^{b} V_i = V$. *Write* γ_B *for the spectral gap of the block dynamics and* γ *for the spectral gap of the single-site dynamics. Then*

$$\gamma_B \le \left[M^2 \cdot M^\star \cdot (4e^{2\beta\Delta})^{M+1}\right] \gamma.$$

PROOF. We will apply the Comparison Theorem (Theorem 13.23), which requires that we define, for each block move from σ to τ, a sequence of single-site moves starting from σ and ending at τ.

For σ and τ which differ only in the block W, define the path $\Gamma_{\sigma,\tau}$ as follows: enumerate the vertices where σ and τ differ as $v_1, \ldots, v_r$. Obtain the k-th state in the path from the $(k-1)$-st by flipping the spin at v_k.

For these paths, we must bound the congestion ratio, defined in (13.20) and denoted here by R.

Suppose that $e = (\sigma_0, \tau_0)$, where σ_0 and τ_0 agree everywhere except at vertex v. Since $\tilde{P}(\sigma, \tau) > 0$ only for σ and τ which differ by a single block update, $|\Gamma_{\sigma,\tau}| \le M$ whenever $\tilde{P}(\sigma, \tau) > 0$. Therefore,

$$R_e := \frac{1}{Q(e)} \sum_{\substack{\sigma,\tau \\ e \in \Gamma_{\sigma,\tau}}} \pi(\sigma)\tilde{P}(\sigma,\tau)|\Gamma_{\sigma,\tau}| \le M \sum_{\substack{\sigma,\tau \\ e \in \Gamma_{\sigma,\tau}}} \frac{1}{b} \sum_{i\,:\, v \in V_i} \frac{\pi(\sigma)P_{V_i}(\sigma,\tau)}{\pi(\sigma_0)P(\sigma_0,\tau_0)}. \tag{15.15}$$

Observe that if σ and τ differ at r vertices, say $D = \{v_1, \ldots, v_r\}$, then

$$\begin{aligned} \frac{\pi(\sigma)}{\pi(\tau)} &= \frac{\exp\left(\beta \sum_{\{u,w\} \cap D \ne \varnothing} \sigma(u)\sigma(w)\right)}{\exp\left(\beta \sum_{\{u,w\} \cap D \ne \varnothing} \tau(u)\tau(w)\right)} \\ &\le e^{2\beta\Delta r}. \end{aligned} \tag{15.16}$$

Write $\sigma \xrightarrow{V_i} \tau$ to indicate that τ can be obtained from σ by a V_i-block update. Bounding $P_{V_i}(\sigma, \tau)$ above by $\mathbf{1}\{\sigma \xrightarrow{V_i} \tau\}$ and $P(\sigma_0, \tau_0)$ below by $1/(2ne^{2\beta\Delta})$ yields

$$\frac{P_{V_i}(\sigma,\tau)}{P(\sigma_0,\tau_0)} \le 2ne^{2\beta\Delta}\mathbf{1}\{\sigma \xrightarrow{V_i} \tau\}. \tag{15.17}$$

Using the bounds (15.16) and (15.17) in (15.15) shows that

$$R_e \le \left(\frac{M}{b}\right)\left(2ne^{2\beta\Delta}\right)\left(e^{2\beta\Delta}\right)^M \sum_i \mathbf{1}\{v \in V_i\} \sum_{\substack{\sigma,\tau \\ e \in \Gamma_{\sigma,\tau}}} \mathbf{1}\{\sigma \xrightarrow{V_i} \tau\}. \tag{15.18}$$

Since configurations σ and τ differing in a V_i-block move and satisfying $e \in \Gamma_{\sigma,\tau}$ both agree with σ_0 outside V_i, there are most $(2^M)^2 = 4^M$ such pairs. Therefore, by (15.18),

$$R := \max_e R_e \le 2\left(\frac{n}{b}\right) M e^{2\beta\Delta(M+1)} M^\star 4^M.$$

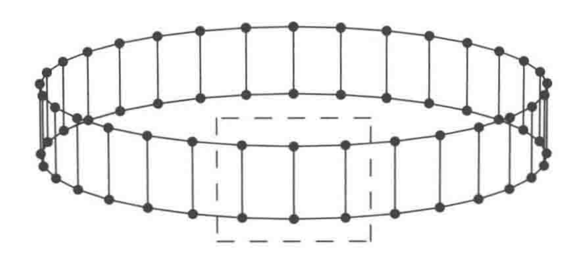


FIGURE 15.3. The ladder graph with $n = 32$. The set of vertices enclosed in the dashed box is a block of length $\ell = 3$.

Since there is at least one block for each site by the hypothesis that $\bigcup V_i = V$, we have $(n/b) \leq M$. Finally, we achieve the bound $B \leq M^2 \cdot M^\star \left(4e^{2\beta\Delta}\right)^{M+1}$. ■

The ladder graph shown in Figure 15.3 is essentially a one-dimensional graph, so in view of Theorem 15.4 it should not be surprising that at any temperature it has a mixing time of the order $n \log n$. The proof is a very nice illustration of the technique of comparing the single-site dynamics to block dynamics.

Write L_n for the ***circular ladder graph*** having vertex set $\mathbb{Z}_n \times \{0,1\}$ and edge set

$$\{\ \{(k,a),\ (j,a)\}\ :\ j \equiv k-1 \mod n,\ a \in \{0,1\}\ \} \cup \{\ \{(k,0),(k,1)\}\ :\ k \in \mathbb{Z}_n\ \}.$$

See Figure 15.3 for an example with $n = 32$. We will call an edge of the form $\{(k,0),(k,1)\}$ a ***rung***.

THEOREM 15.10. *Let L_n denote the circular ladder graph defined above. There exist $c_0(\beta)$ and $c_1(\beta)$, not depending on n, such that the Glauber dynamics for the Ising model on L_n satisfies $t_{\text{rel}} \leq c_0(\beta)n$, whence $t_{\text{mix}} \leq c_1(\beta)n^2$.*

PROOF. Define the random variable Υ_k on the probability space $(\{-1,1\}^V, \pi)$ by $\Upsilon_k(\sigma) := (\sigma(k,0), \sigma(k,1))$. That is, $\Upsilon_k(\sigma)$ is the pair of spins on the k-th rung in configuration σ.

Define the ***j-th ℓ-block*** to be the vertex set

$$V_j := \{(k,a)\ :\ j+1 \leq k \leq j+\ell,\ a \in \{0,1\}\}.$$

For $j \leq i < j+\ell$, the conditional distribution of Υ_{i+1}, given $(\Upsilon_j, \dots, \Upsilon_i)$ and $\Upsilon_{j+\ell+1}$, depends only on Υ_i and $\Upsilon_{j+\ell+1}$. Therefore, given Υ_j and $\Upsilon_{j+\ell+1}$, the sequence $(\Upsilon_i)_{i=j}^{j+\ell}$ is a time-inhomogeneous Markov chain. If block V_j is selected to be updated in the block dynamics, the update can be realized by running this chain. We call this the ***sequential*** method of updating.

We now describe how to couple the block dynamics started from σ with the block dynamics started from τ, in the case that σ and τ differ at only a single site, say (j,a). Always select the same block to update in both chains. If a block is selected which contains (j,a), then the two chains can be updated together, and the difference at (j,a) is eliminated. The only difficulty occurs when (j,a) is a neighbor of a vertex belonging to the selected block.

We treat the case where block V_j is selected; the case where the block is immediately to the left of (j,a) is identical. We will use the sequential method of updating on both chains. Let $(\Upsilon_i)_{i=j}^{j+\ell}$ denote the chain used to update σ, and let

$(\tilde{\Upsilon}_i)_{i=j}^{j+\ell}$ denote the chain used to update τ. We run Υ and $\tilde{\Upsilon}$ independently until they meet, and after the two chains meet, we perform identical transitions in the two chains.

Since $\pi(\sigma)/\pi(\tilde{\sigma}) \leq e^{12\beta}$ when σ and $\tilde{\sigma}$ differ on a rung, the probability that the spins on a rung take any of the four possible ± 1 pairs, given the spins outside the rung, is at least $[4e^{12\beta}]^{-1}$. Thus, as the sequential update chains move across the rungs, at each rung there is a chance of at least $(1/4)e^{-24\beta}$, given the previous rungs, that the two chains will have the same value. Therefore, the expected total number of vertices where the two updates disagree is bounded by $8e^{24\beta}$.

Let ρ denote Hamming distance between configurations, so $\rho(\sigma, \tau) = 1$. Let (X_1, Y_1) be the pair of configurations obtained after one step of the coupling. Since ℓ of the n blocks will contain (j, a) and two of the blocks have vertices neighboring (j, a), we have

$$\mathbf{E}_{\sigma,\tau}\rho(X_1, Y_1) \leq 1 - \frac{\ell}{n} + \frac{2}{n}8e^{24\beta}.$$

If we take $\ell = \ell(\beta) = 16e^{24\beta} + 1$, then

$$\mathbf{E}_{\sigma,\tau}\rho(X_1, Y_1) \leq 1 - \frac{1}{n} \leq e^{-1/n} \tag{15.19}$$

for any σ and τ with $\rho(\sigma, \tau) = 1$. By Theorem 14.6, for any two configurations σ and τ, there exists a coupling (X_1, Y_1) of the block dynamics satisfying

$$\mathbf{E}_{\sigma,\tau}\rho(X_1, Y_1) \leq \rho(\sigma, \tau)e^{-1/n}.$$

Theorem 13.1 implies that $\gamma_B \geq 1/n$. By Theorem 15.9, we conclude that $\gamma \geq c_0(\beta)/n$ for some $c_0(\beta) > 0$. Applying Theorem 12.3 shows that $t_{\text{mix}} \leq c_1(\beta)n^2$. ∎

REMARK 15.11. In fact, for the Ising model on the circular ladder graph, $t_{\text{mix}} \leq c(\beta)n \log n$, although different methods are needed to prove this. See Martinelli (1999).

15.6. Lower Bound for Ising on Square*

Consider the Glauber dynamics for the Ising model in an $n \times n$ box: $V = \{(j,k) : 0 \leq j, k \leq n-1\}$ and edges connect vertices at unit Euclidean distance.

In this section we prove

THEOREM 15.12 (Schonmann (1987) and Thomas (1989)). *The relaxation time $(1-\lambda_\star)^{-1}$ of the Glauber dynamics for the Ising model in an $n \times n$ square in two dimensions is at least $\exp(\psi(\beta)n)$, where $\psi(\beta) > 0$ if β is large enough.*

More precisely, let $\gamma_\ell < 3^\ell$ be the number of self-avoiding lattice paths starting from the origin in $\mathbb{Z}^2$ that have length ℓ, and let $\gamma < 3$ be the "connective constant" for the planar square lattice, defined by $\gamma := \lim_{\ell\to\infty} \sqrt[\ell]{\gamma_\ell}$. If $\beta > (1/2)\log(\gamma)$, then $\psi(\beta) > 0$.

Much sharper and more general results are known; see the partial history in the notes. We provide here a proof following closely the method used by Dana Randall (2006) for the hardcore lattice gas.

The key idea in Randall (2006) is not to use the usual cut determined by the magnetization (as in the proof of Theorem 15.3), but rather a topological obstruction. As noted by Fabio Martinelli (personal communication), this idea was already present in Thomas (1989), where contours were directly used to define a cut and obtain the right order lower bound for the relaxation time. Thus the present discussion

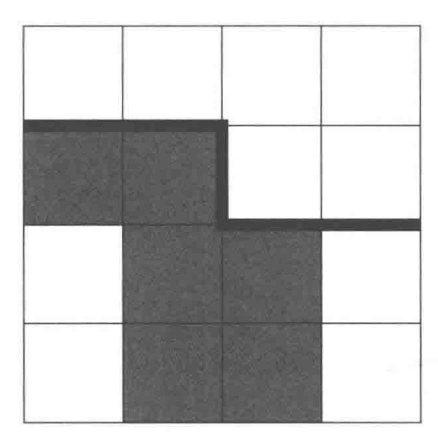


FIGURE 15.4. A fault line with one defect. Positive spins are indicated by shaded squares, while negative spins are indicated by white squares. The fault line is drawn in bold.

is purely expository with no claim of originality. The argument in Thomas (1989) works in all dimensions and hence is harder to read.

REMARK 15.13. An upper bound on relaxation time of order $\exp(C(\beta)n^{d-1})$ in all dimensions follows from the "path method" of Jerrum and Sinclair (1989) for all β. The constant $C(\beta)$ obtained that way is not optimal.

In proving Theorem 15.12, it will be convenient to attach the spins to the faces (lattice squares) of the lattice rather than the nodes.

DEFINITION 15.14. A ***fault line*** (with at most k defects) is a self-avoiding lattice path from the left side to the right side or from the top to the bottom of $[0,n]^2$, where each edge of the path (with at most k exceptions) is adjacent to two faces with different spins on them. Thus no edges in the fault line are on the boundary of $[0,n]^2$. See Figure 15.4 for an illustration.

LEMMA 15.15. *Denote by F_k the set of Ising configurations in $[0,n]^2$ that have a fault line with at most k defects. Then $\pi(F_k) \le \sum_{\ell \ge n} 2\ell\gamma_\ell e^{2\beta(2k-\ell)}$. In particular, if k is fixed and $\beta > (1/2)\log(\gamma)$, then $\pi(F_k)$ decays exponentially in n.*

PROOF. For a self-avoiding lattice path φ of length ℓ from the left side to the right side (or from top to bottom) of $[0,n]^2$, let F_φ be the set of Ising configurations in $[0,n]^2$ that have φ as a fault line with at most k defects. Reflecting all the spins on one side of the fault line (say, the side that contains the upper left corner) defines a one-to-one mapping from F_φ to its complement that magnifies probability by a factor of $e^{2\beta(\ell-2k)}$. This yields that $\pi(F_\varphi) \le e^{2\beta(2k-\ell)}$.

Summing this over all self-avoiding lattice paths φ of length ℓ from top to bottom and from left to right of $[0,n]^2$ and over all $\ell \ge n$ completes the proof. ■

LEMMA 15.16.

(i) *If in a configuration σ there is no all-plus crossing from the left side L of $[0,n]^2$ to the right side R and there is also no all-minus crossing, then there is a fault line with no defects from the top to the bottom of $[0,n]^2$.*

(ii) *Similarly, if* Γ_+ *is a path of lattice squares (all labeled plus in* σ*) from a square* q *in* $[0,n]^2$ *to the top side of* $[0,n]^2$ *and* Γ_- *is a path of lattice squares (all labeled minus) from the same square* q *to the top of* $[0,n]^2$*, then there is a lattice path* ξ *from the boundary of* q *to the top of* $[0,n]^2$ *such that every edge in* ξ *is adjacent to two lattice squares with different labels in* σ*.*

PROOF.

(i) For the first statement, let A be the collection of lattice squares that can be reached from L by a path of lattice squares of the same label in σ. Let $A^\star$ denote the set of squares that are separated from R by A. Then the boundary of $A^\star$ consists of part of the boundary of $[0,n]^2$ and a fault line.

(ii) Suppose q itself is labeled minus in σ and Γ_+ terminates in a square q_+ on the top of $[0,n]^2$ which is to the left of the square q_- where Γ_- terminates. Let A_+ be the collection of lattice squares that can be reached from Γ_+ by a path of lattice squares labeled plus in σ and denote by $A_+^\star$ the set of squares that are separated from the boundary of $[0,n]^2$ by A_+. Let ξ_1 be a directed lattice edge with q on its right and a square of Γ_+ on its left. Continue ξ_1 to a directed lattice path ξ leading to the boundary of $[0,n]^2$, by inductively choosing the next edge ξ_j to have a square (labeled plus) of A_+ on its left and a square (labeled minus) not in $A_+^\star$ on its right. It is easy to check that such a choice is always possible (until ξ reaches the boundary of $[0,n]^2$]), the path ξ cannot cycle and it must terminate between q_+ and q_- on the top side of $[0,n]^2$.

■

PROOF OF THEOREM 15.12. Following Randall (2006), let S_+ be the set of configurations that have a top-to-bottom and a left-to-right crossing of pluses. Similarly define S_-. On the complement of $S_+ \cup S_-$ there is either no monochromatic crossing left-to-right (whence there is a top-to-bottom fault line by Lemma 15.16) or there is no monochromatic crossing top-to-bottom (whence there is a left-to-right fault line). By Lemma 15.15, $\pi(S_+) \to 1/2$ as $n \to \infty$.

Let ∂S_+ denote the external vertex boundary of S_+, that is, the set of configurations outside S_+ that are one flip away from S_+. It suffices to show that $\pi(\partial S_+)$ decays exponentially in n for $\beta > \frac{1}{2}\log(\gamma)$. By Lemma 15.15, it is enough to verify that every configuration $\sigma \in \partial S_+$ has a fault line with at most 3 defects.

The case $\sigma \notin S_-$ is handled by Lemma 15.16. Fix $\sigma \in \partial S_+ \cap S_-$ and let q be a lattice square such that flipping $\sigma(q)$ will transform σ to an element of S_+. By Lemma 15.16, there is a lattice path ξ from the boundary of q to the top of $[0,n]^2$ such that every edge in ξ is adjacent to two lattice squares with different labels in σ; by symmetry, there is also such a path $\xi^\star$ from the boundary of q to the bottom of $[0,n]^2$. By adding at most three edges of q, we can concatenate these paths to obtain a fault line with at most three defects.

Lemma 15.15 completes the proof. ■

Exercises

EXERCISE 15.1. Let (G_n) be a sequence of expander graphs with maximal degree Δ. Find $\beta(\Delta)$ such that for $\beta > \beta(\Delta)$, the relaxation time for Glauber dynamics for the Ising model grows exponentially in n.

EXERCISE 15.2. Consider the Ising model on the b-ary tree of depth k, and let $f(\sigma) = \sum_{v\,:\,|v|=k} \sigma(v)$. Let $\theta = \tanh(\beta)$. Show that

$$\mathrm{Var}_\pi(f) \asymp \sum_{j=0}^{k} b^{k+j}\theta^{2j} \asymp \begin{cases} b^k & \text{if } \theta < 1/\sqrt{b}, \\ kb^k \asymp n \log n & \text{if } \theta = 1/\sqrt{b}, \\ (b\theta)^{2k}) \asymp n^{1+\alpha} & \text{if } \theta > 1/\sqrt{b}, \end{cases}$$

where $\alpha = \log(b\theta^2)/\log(b) > 0$. Use this to obtain lower bounds on t_{rel} in the three regimes.

Notes

The upper and lower bounds obtained in Theorem 15.4 for the mixing time for Glauber dynamics on the cycle are within a factor of two of each other. An enticing open problem is to show that one of these two bounds is sharp. This chain is an example where pre-cutoff is known, although cutoff has not been proven. (See Chapter 18 for the definitions of pre-cutoff and cutoff.)

Kenyon, Mossel, and Peres (2001) showed that the relaxation time of the Glauber dynamics for the Ising model on the b-ary tree has the following behavior: if $\theta < 1/\sqrt{b}$, then $t_{\mathrm{rel}} = O(n)$, while if $\theta > 1/\sqrt{b}$, then $t_{\mathrm{rel}} \geq c_1 n^{1+\alpha}$, where $\alpha > 0$ depends on β. The case $\theta > 1/\sqrt{b}$ can be proved by using the function $f(\sigma) = \sum_{\text{leaves}} \sigma(v)$ in the variational principle (Lemma 13.12); see Exercise 15.2. See Berger, Kenyon, Mossel, and Peres (2005) and Martinelli, Sinclair, and Weitz (2004) for extensions.

Theorem 15.3 does not say what happens when $\beta = 1$. Levin, Luczak, and Peres (2007) showed that at $\beta = 1$, the mixing time is $O(n^{3/2})$.

Levin, Luczak, and Peres (2007) also showed that if $\beta > 1$ and the dynamics are restricted to the part of the state space where $\sum \sigma(v) > 0$, then $t_{\mathrm{mix}} = O(n \log n)$. In the case where $\beta < 1$, they show that the chain has a cutoff. These results were further refined by Ding, Lubetzky, and Peres (2008a).

A partial history of Ising on the square lattice. For the ferromagnetic Ising model with no external field and free boundary, Schonmann (1987) proved

THEOREM 15.17. *In dimension 2, let $m^\star$ denote the "spontaneous magnetization", i.e., the expected spin at the origin in the plus measure in the whole lattice. Denote by $p(n; a, b)$ the probability that the magnetization (average of spins) in an $n \times n$ square is in an interval (a, b). If $-m^\star < a < b < m^\star$, then $p(n; a, b)$ decays exponentially in n.*

(The rate function was not obtained, only upper and lower bounds.)

Using the easy direction of the Cheeger inequality (Theorem 13.14), which is an immediate consequence of the variational formula for eigenvalues, this yields Theorem 15.12.

Chayes, Chayes and Schonmann (1987) then extended Theorem 15.17 to all $\beta > \beta_c$. (Recall that for the planar square lattice $\beta_c = \log(1 + \sqrt{2})/2$.)

Theorem 15.12 was stated explicitly and proved in Thomas (1989) who extended it to all dimensions $d \geq 2$. He did not use the magnetization to define a cut, but instead his cut was defined by configurations where there is a contour of length (or in higher dimensions $d \geq 3$, surface area) larger than an^{d-1} for a suitable small

$a > 0$. Again the rate function was only obtained up to a constant factor and he assumed β was large enough for a Peierls argument to work.

In the breakthrough book of Dobrushin, Kotecký and Shlosman (1992) the correct rate function (involving surface tension) for the large deviations of magnetization in 2 dimensions was identified and established for large β.

This was extended by Ioffe (1995) to all $\beta > \beta_c$. The consequences for mixing time (a sharp lower bound) and a corresponding sharp upper bound were established in Cesi, Guadagni, Martinelli, and Schonmann (1996).

In higher dimensions, a lower bound for mixing time of the right order (exponential in n^{d-1}) for all $\beta > \beta_c(d, slab)$ follows from the magnetization large deviation bounds of Pisztora (1996). That $\beta_c(d, slab)$ coincides with β_c was proven by Bodineau (2005).

The correct rate function has not yet been determined but a related result under plus boundary conditions is in Cerf and Pisztora (2000).

For more precise results for the Ising model on the lattice and their applications to Glauber dynamics, see Dobrushin and Shlosman (1987), Stroock and Zegarliński (1992), Martinelli and Olivieri (1994), and Martinelli, Olivieri, and Schonmann (1994).

Further reading. An excellent source on dynamics for the Ising model is Martinelli (1999). Simon (1993) contains more on the Ising model. Ising's thesis (published as Ising (1925)) concerned the one-dimensional model. For information on the life of Ising, see Kobe (1997).

CHAPTER 16

From Shuffling Cards to Shuffling Genes

One reasonable restriction of the random transposition shuffle is to only allow interchanges of adjacent cards—see Figure 16.1. Restricting the moves in this

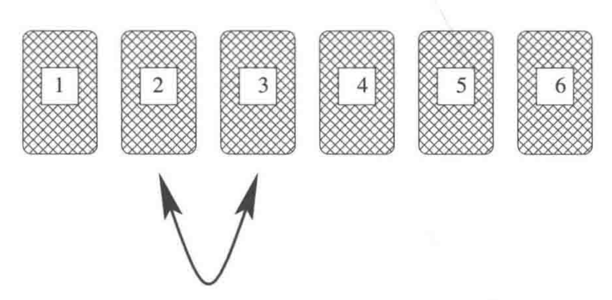


FIGURE 16.1. An adjacent transposition swaps two neighboring cards.

manner slows the shuffle down. It also breaks the symmetry of the random transpositions walk enough to require different methods of analysis.

In Section 16.1 we examine the mixing of the random adjacent transpositions walk using several different methods: upper bounds via comparison (way off) and coupling (quite sharp) and lower bounds via following a single card (off by a log factor) and Wilson's method (sharp).

A generalization of the random adjacent transpositions model, in which entire segments of a permutation are reversed in place, can be interpreted as modeling large-scale genome changes. Varying the maximum allowed length of the reversed segments impacts the mixing time significantly. We study these reversal chains in Section 16.2.

16.1. Random Adjacent Transpositions

As usual we consider a lazy version of the chain to avoid periodicity problems. The resulting increment distribution assigns probability $1/[2(n-1)]$ to each of the transpositions $(1\,2), \ldots, (n-1\,n)$ and probability $1/2$ to id.

16.1.1. Upper bound via comparison. We can bound the convergence of the random adjacent transposition shuffle by comparing it with the random transposition shuffle. While our analysis considers only the spectral gap and thus gives a poor upper bound on the mixing time, we illustrate the method because it can be used for many types of shuffle chains.

Note: in the course of this proof, we will introduce several constants $C_1, C_2, \ldots$. Since we are deriving such (asymptotically) poor bounds, we will not make any effort to optimize their values. Each one does not depend on n.

First, we bound the relaxation time of the random transpositions shuffle by its mixing time. By Theorem 12.4 and Corollary 8.10,

$$t_{\text{rel}} = O(n \log n). \tag{16.1}$$

(We are already off by a factor of $\log n$, but we will lose so much more along the way that it scarcely matters.)

Next we compare. In order to apply Corollary 13.27, we must express an arbitrary transposition $(a\,b)$, where $1 \le a < b \le n$, in terms of adjacent transpositions. Note that

$$(a\,b) = (a\,a+1)\dots(b-1\,b-2)(b-1\,b)(b-1\,b-2)\dots(a+1\,a+2)(a\,a+1). \tag{16.2}$$

This path has length at most $2n-3$ and uses any single adjacent transposition at most twice.

We must estimate the congestion ratio

$$B = \max_{s\in S} \frac{1}{\mu(s)} \sum_{a \in \tilde{S}} \tilde{\mu}(a) N(s,a)\,|a| \le \max_{s \in S} \frac{4(n-1)}{n^2} \sum_{a\in\tilde{S}} N(s,a)\,|a|. \tag{16.3}$$

Here S is the support of the random adjacent transposition walk, μ is its increment distribution, $\tilde{S}$ and $\tilde{\mu}$ are the corresponding features of the random transpositions walk, $N(s,a)$ is the number of times s is used in the expansion of a, and $|a|$ is the total length of the expansion of a. Since an adjacent transposition $s = (i\,i+1)$ lies on the generator path of $(a\,b)$ exactly when $a \le i < i+1 \le b$, no generator path uses any adjacent transposition more than twice, and the length of the generator paths is bounded by $(2n-3)$, the summation on the right-hand-side of (16.3) is bounded by $2i(n-i)(2n-3) \le n^3$. Hence

$$B \le 4n^2,$$

and Corollary 13.27 tells us that the relaxation time of the random adjacent transpositions chain is at most $C_2 n^3 \log n$.

Finally, we use Theorem 12.3 to bound the mixing time by the relaxation time. Here the stationary distribution is uniform, $\pi(\sigma) = 1/n!$ for all $\sigma \in \mathcal{S}_n$. The mixing time of the random adjacent transpositions chain thus satisfies

$$t_{\text{mix}} \le \log(4n!) C_2 n^3 \log n \le C_3 n^4 \log^2 n.$$

16.1.2. Upper bound via coupling. The coupling we present here is described in Aldous (1983b) and also discussed in Wilson (2004a).

In order to couple two copies (σ_t) and (σ'_t) (the "left" and "right" decks) of the lazy version or the random adjacent transpositions chain, proceed as follows. First, choose a pair $(i, i+1)$ of adjacent locations uniformly from the possibilities. Flip a coin to decide whether to perform the transposition on the left deck. Now, examine the cards $\sigma_t(i), \sigma'_t(i), \sigma_t(i+1)$ and $\sigma'_t(i+1)$ in locations i and $i+1$ in the two decks.

- If $\sigma_t(i) = \sigma'_t(i+1)$ or if $\sigma_t(i+1) = \sigma'_t(i)$, then do the opposite to the right deck: transpose if the left deck stayed still, and vice versa.
- Otherwise, perform the same action on the right deck as on the left deck.

We consider first τ_a, the time required for a particular card a to reach the same position in the two decks. Let X_t be the (unsigned) distance between the positions of a in the two decks at time t. Our coupling ensures that $|X_{t+1} - X_t| \le 1$ and that if $t \ge \tau_a$, then $X_t = 0$.

Let M be the transition matrix of a random walk on the path with vertices $\{0, \dots, n-1\}$ that moves up or down, each with probability $1/(n-1)$, at all interior vertices; from $n-1$ it moves down with probability $1/(n-1)$, and, under all other circumstances, it stays where it is. In particular, it absorbs at state 0.

Note that for $1 \le i \le n-1$,

$$\mathbf{P}\{X_{t+1} = i-1 \mid X_t = i, \sigma_t, \sigma_t'\} = M(i, i-1).$$

However, since one or both of the cards might be at the top or bottom of a deck and thus block the distance from increasing, we can only say

$$\mathbf{P}\{X_{t+1} = i+1 \mid X_t = i, \sigma_t, \sigma_t'\} \le M(i, i+1).$$

Even though the sequence (X_t) is not a Markov chain, the above inequalities imply that we can couple it to a random walk (Y_t) with transition matrix M in such a way that $Y_0 = X_0$ and $X_t \le Y_t$ for all $t \ge 0$. Under this coupling τ_a is bounded by the time τ_0^Y it takes (Y_t) to absorb at 0.

The chain (Y_t) is best viewed as a delayed version of a simple random walk on the path $\{0, \dots, n-1\}$, with a hold probability of $1/2$ at $n-1$ and absorption at 0. At interior nodes, with probability $1 - 2/(n-1)$, the chain (Y_t) does nothing, and with probability $2/(n-1)$, it takes a step in that walk. Exercises 2.3 and 2.2 imply that $\mathbf{E}(\tau_0^Y)$ is bounded by $(n-1)n^2/2$, regardless of initial state. Hence

$$\mathbf{E}(\tau_a) < \frac{(n-1)n^2}{2}.$$

By Markov's inequality,

$$\mathbf{P}\{\tau_a > n^3\} < 1/2$$

for sufficiently large n. If we run $2\log_2 n$ blocks, each consisting of n^3 shuffles, we can see that

$$\mathbf{P}\{\tau_a > 2n^3 \log_2 n\} < \frac{1}{n^2}.$$

Now let's look at all the cards. After $2n^3 \log_2 n$ steps, the probability of the decks having not coupled is bounded by the sum of the probabilities of the individual cards having not coupled, so

$$\mathbf{P}\{\tau_{\text{couple}} > 2n^3 \log_2 n\} < \frac{1}{n}, \tag{16.4}$$

regardless of the initial states of the decks. Theorem 5.2 immediately implies that $t_{\text{mix}}(\varepsilon) < 2n^3 \log_2 n$ for sufficiently large n.

16.1.3. Lower bound via following a single card. Consider the set of permutations

$$A = \{\sigma : \sigma(1) \ge \lfloor n/2 \rfloor\}.$$

Under the uniform distribution we have $U(A) = (n - (\lfloor n/2 \rfloor - 1))/n \ge 1/2$, because card 1 is equally likely to be in any of the n possible positions. However, since card 1 can change its location by at most one place in a single shuffle and since card 1 does not get to move very often, it is plausible that a large number of shuffles must be applied to a sorted deck before the event A has reasonably large probability. Below we formalize this argument.

How does card 1 move under the action of the random adjacent transposition shuffle? Let us first make the general observation that when (σ_t) is a random walk on $\mathcal{S}_n$ with increment distribution Q and $k \in [n]$, Lemma 2.5 implies that

the sequence $(\sigma_t(k))$ is itself a Markov chain, which we will call the ***single-card chain***. Its transition matrix P' does not depend on k.

Returning to the case of (lazy) random adjacent transpositions: each interior card (neither top nor bottom of the deck) moves with probability $1/(n-1)$, and at each of the moves it is equally likely to jump one position to the right or one position to the left. If the card is at an endpoint, it is selected with probability $1/2(n-1)$ and always moves in the one permitted direction. If $(\tilde{S}_t)$ is a random walk on $\mathbb{Z}$ which remains in place with probability $1-1/(n-1)$ and increments by ± 1 with equal probability when it moves, then

$$\mathbf{P}\{\sigma_t(1) - 1 \geq z\} \geq \mathbf{P}\{|\tilde{S}_t| \geq z\}.$$

Thus,

$$\mathbf{P}\{\sigma_t(1) \geq n/2 + 1\} \leq \frac{4\mathbf{E}\tilde{S}_t^2}{n^2} \leq \frac{4t}{n^2(n-1)}.$$

Therefore,

$$\left\|P^t(\mathrm{id}, \cdot) - U\right\|_{\mathrm{TV}} \geq U(A) - P^t(\mathrm{id}, A) \geq \frac{1}{2} - \frac{4t}{n^2(n-1)}.$$

Thus if $t \leq n^2(n-1)/16$, then $d(t) \geq 1/4$. We conclude that $t_{\mathrm{mix}} \geq n^2(n-1)/16$.

16.1.4. Lower bound via Wilson's method. In order to apply Wilson's method (Theorem 13.5) to the random adjacent transpositions shuffle, we must specify an eigenfunction and initial state.

First, some generalities on the relationship between the eigenvalues and eigenfunctions of a shuffle chain and its single-card chain. Lemma 12.8 tells us that when $\Phi : [n] \to \mathbb{R}$ is an eigenfunction of the single-card chain with eigenvalue λ, then $\Phi^\flat : \mathcal{S}_n \to \mathbb{R}$ defined by $\Phi^\flat(\sigma) = \Phi(\sigma(k))$ is an eigenfunction of the shuffle chain with eigenvalue λ.

For the random adjacent transpositions chain, the single-card chain is an extremely lazy version of a random walk on the path whose eigenfunctions and eigenvalues were determined in Section 12.3.2. Let M be the transition matrix of simple random walk on the n-path with holding probability $1/2$ at the endpoints. Then we have

$$P' = \frac{1}{n-1}M + \frac{n-2}{n-1}I.$$

It follows from (12.18) that

$$\varphi(k) = \cos\left(\frac{(2k-1)\pi}{2n}\right)$$

is an eigenfunction of P' with eigenvalue

$$\lambda = \frac{1}{n-1}\cos\left(\frac{\pi}{n}\right) + \frac{n-2}{n-1} = 1 - \frac{\pi^2}{2n^3} + O\left(\frac{1}{n^3}\right).$$

Hence, for any $k \in [n]$ the function $\sigma \mapsto \varphi(\sigma(k))$ is an eigenfunction of the random transposition walk with eigenvalue λ. Since these eigenfunctions all lie in the same eigenspace, so will any linear combination of them. We set

$$\Phi(\sigma) = \sum_{k\in[n]} \varphi(k)\varphi(\sigma(k)). \tag{16.5}$$

REMARK 16.1. See Exercise 8.9 for some motivation of our choice of Φ. By making sure that $\Phi(\mathrm{id})$ is as large as possible, we ensure that when $\Phi(\sigma_t)$ is small, then σ_t is in some sense likely to be far away from the identity.

Now consider the effect of a single adjacent transposition $(k-1\,k)$ on Φ. Only two terms in (16.5) change, and we compute

$$\begin{aligned}|\Phi(\sigma(k-1\,k)) - \Phi(\sigma)| &= |\varphi(k)\varphi(\sigma(k-1)) + \varphi(k-1)\varphi(\sigma(k))\\ &\quad - \varphi(k-1)\varphi(\sigma(k-1)) - \varphi(k)\varphi(\sigma(k))|\\ &= |(\varphi(k)-\varphi(k-1))(\varphi(\sigma(k)) - \varphi(\sigma(k-1)))|.\end{aligned}$$

Since $d\varphi(x)/dx$ is bounded in absolute value by π/n and $\varphi(x)$ itself is bounded in absolute value by 1, we may conclude that

$$|\Phi(\sigma(k-1\,k)) - \Phi(\sigma)| \le \frac{\pi}{n}(2) = \frac{2\pi}{n}. \tag{16.6}$$

Combining (16.6) with Theorem 13.5 and the fact that $\Phi(\mathrm{id}) = n/2$ (see Exercise 8.10) tells us that when the random adjacent transposition shuffle is started with a sorted deck, after

$$t = \frac{n^3 \log n}{\pi^2} + C_\varepsilon n^3 \tag{16.7}$$

steps the variation distance from stationarity is still at least ε. (Here C_ε can be taken to be $\log\left(\frac{1-\varepsilon}{64\varepsilon}\right)$.)

16.2. Shuffling Genes

Although it is amusing to view permutations as arrangements of a deck of cards, they occur in many other contexts. For example, there are (rare) mutation events involving large-scale rearrangements of segments of DNA. Biologists can use the relative order of homologous genes to estimate the evolutionary distance between two organisms. Durrett (2003) has studied the mixing behavior of the random walk on $\mathcal{S}_n$ corresponding to one of these large-scale rearrangement mechanisms, *reversals*.

Fix $n > 0$. For $1 \le i \le j \le n$, define the ***reversal*** $\rho_{i,j} \in \mathcal{S}_n$ to be the permutation that reverses the order of all elements in places i through j. (The reversal $\rho_{i,i}$ is simply the identity.)

Since not all possible reversals are equally likely in the chromosomal context, we would like to be able to limit what reversals are allowed as steps in our random walks. One (simplistic) restrictive assumption is to require that the endpoints of the reversal are at distance at most L from each other.

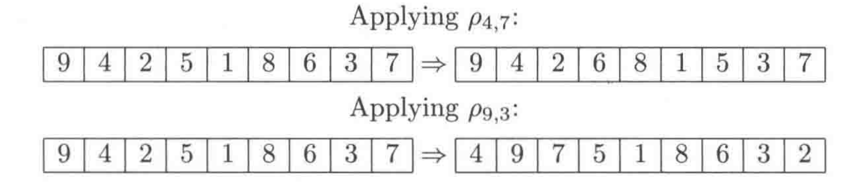


FIGURE 16.2. Applying reversals to permutations of length 9. Note that the second reversal wraps around the ends of the permutation.

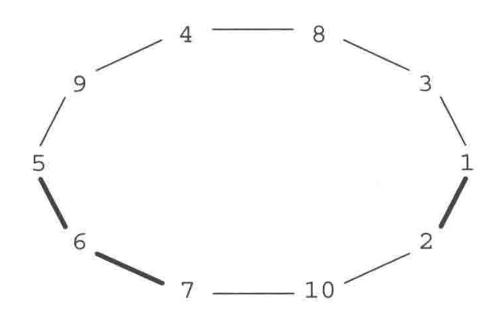


FIGURE 16.3. The permutation 1, 3, 8, 4, 9, 5, 6, 7, 10, 2 has three conserved edges.

To avoid complications at the ends of segments, we will treat our sequences as circular arrangements. Reversals will be allowed to span the "join" in the circle, and all positions will be treated mod n. See Figure 16.2. With these assumptions, we are now ready to define the L-reversal walk.

Let $L = L(n)$ be a function of n satisfying $1 \leq L(n) \leq n$. The ***L-reversal chain*** on $\mathcal{S}_n$ is the random walk on $\mathcal{S}_n$ whose increment distribution is uniform on the set of all reversals of (circular) segments of length at most L. (Note that this includes the n segments of length 1; reversing a segment of length 1 results in the identity permutation.)

Equivalently, to perform a step in the L-reversal chain: choose $i \in [n]$ uniformly, and then choose $k \in [0, L-1]$ uniformly. Perform the reversal $\rho_{i,i+k}$ (which will wrap around the ends of the sequence when $i + k > n$). Note that the total probability assigned to id is $n/nL = 1/L$.

Since each reversal is its own inverse, Proposition 2.14 ensures that the L-reversal chain is reversible.

In Section 16.2.1 we give a lower bound on the mixing time of the L-reversal chain that is sharp in some cases. In Section 16.2.2, we will present an upper bound for their mixing.

16.2.1. Lower bound. Although a single reversal can move many elements, it can break at most two adjacencies. We use the number of preserved adjacencies to lower bound the mixing time.

PROPOSITION 16.2. *Consider the family of L-reversal chains, where $L = L(n)$ satisfies $1 \leq L(n) < n/2$. Fix $0 < \varepsilon < 1$ and let $t = t(n) = (1-\varepsilon)\frac{n}{2}\log n$. Then*

$$\lim_{n\to\infty} d(t) = 1.$$

PROOF. Superimpose the edges of a cycle onto our permutation, and say an edge is ***conserved*** if its endpoints are consecutive—in either order (see Figure 16.3).

Under the uniform distribution on $\mathcal{S}_n$, each cycle edge has probability $2/n$ of being conserved. Hence the expected number of conserved edges is 2.

Now consider running the L-reversal chain. Each reversal breaks the cycle at 2 edges and reverses the segment in between them. Call an edge ***undisturbed*** if it has not been cut by any reversal. There are two reasons that a disturbed edge might end up conserved: a reversal of a segment of length 1 is simply the identity permutation and does not change adjacencies, and vertices cut apart by one reversal

might be moved back together by a later one. However, after t reversals, we may be sure that the number of conserved edges is at least as large as the number of undisturbed edges.

Start running the L-reversal chain from the identity permutation, and let U be the number of undisturbed edges at time $t = t(n)$. We can write $U = U_1 + \cdots + U_n$, where U_k is the indicator of the edge $(k, k+1)$ being undisturbed. Under the L-reversal model, each edge has probability $2/n$ of being disturbed in each step, so

$$\mathbf{E}U = n\left(1 - \frac{2}{n}\right)^t \sim n^{\varepsilon}.$$

We can also use indicators to estimate the variance of U. At each step of the chain, there are nL reversals that can be chosen. Each edge is disturbed by exactly $2L$ legal reversals, since it can be either the right or the left endpoint of reversals of L different lengths. If the edges are more than L steps apart, no legal reversal breaks both. If they are closer than that, exactly one reversal breaks both. Hence

$$\mathbf{P}\{U_i = 1 \text{ and } U_j = 1\} = \begin{cases} \left(\frac{nL-(4L-1)}{nL}\right)^t & \text{if } 1 \le j - i \le L \text{ or } 1 \le i - j \le L, \\ \left(\frac{nL-4L}{nL}\right)^t & \text{otherwise} \end{cases}$$

(in this computation, the subscripts must be interpreted mod n).

Write $p = \mathbf{P}(U_k = 1) = (1 - 2/n)^t \sim n^{\varepsilon - 1}$. We can now estimate

$$\begin{aligned} \mathrm{Var}U &= \sum_{i=1}^{n} \mathrm{Var}U_i + \sum_{i \neq j} \mathrm{Cov}(U_i, U_j) \\ &= np(1-p) + 2nL\left(\left(1 - \frac{4 - 1/L}{n}\right)^t - p^2\right) \\ &\quad + n(n - 2L)\left(\left(1 - \frac{4}{n}\right)^t - p^2\right). \end{aligned}$$

Note that the sum of covariances has been split into those terms for which i and j are at a distance at most L apart and those for which they are further apart. Let's examine the resulting pieces individually. For the first one, factoring out p^2 and taking a power series expansion gives

$$p^2 \cdot 2nL\left(\left(1 + \frac{1}{nL} + O\left(\frac{1}{n^2}\right)\right)^t - 1\right) = O\left(\frac{p^2 nLt}{nL}\right) = o(np),$$

so these terms (which are positive) are negligible compared to $\mathbf{E}U$.

Doing the same to the second piece yields

$$n(n-k)p^2\left(\left(1 - \frac{4}{n^2 - 4n + 4}\right)^t - 1\right) = O\left(n^2 p^2 \cdot \frac{t}{n^2}\right) = o(np),$$

so that these terms (which are negative) are also negligible compared to $\mathbf{E}U$. Since $p = o(1)$, we can conclude that

$$\mathrm{Var}U \sim \mathbf{E}U. \tag{16.8}$$

Let $A \subseteq \mathcal{S}_n$ be the set of permutations with at least $\mathbf{E}U/2$ conserved edges. Under the uniform distribution on $\mathcal{S}_n$, the event A has probability less than or equal to $4/\mathbf{E}U$, by Markov's ineqality.

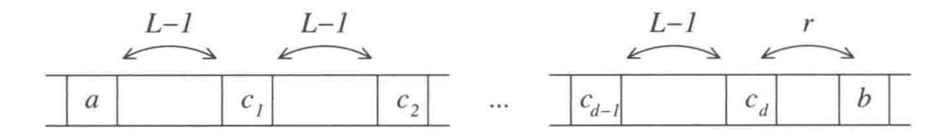


FIGURE 16.4. To express $(a\,b)$ in terms of short transpositions, first carry the marker at position a over to position b; then perform all but the last transposition in reverse order to take the marker at position b over to position a.

By Chebyshev's inequality and (16.8), for sufficiently large n we have

$$P^t(\mathrm{id}, A^c) \le \mathbf{P}\{|U - \mathbf{E}U| > \mathbf{E}U/2\} \le \frac{\mathrm{Var}U}{(\mathbf{E}U/2)^2} < \frac{5}{\mathbf{E}U}.$$

By the definition (4.1) of total variation distance,

$$\left\|P^t(\mathrm{id}, \cdot) - U\right\|_{\mathrm{TV}} \ge \left(1 - \frac{5}{\mathbf{E}U}\right) - \frac{4}{\mathbf{E}U} = 1 - \frac{9}{\mathbf{E}U}.$$

Since $\mathbf{E}U \sim n^{\varepsilon}$, we are done. ■

16.2.2. Upper bound. We now give an upper bound on the mixing time of the L-reversal chain via the comparison method, using the same inefficient methods as we did for random adjacent transpositions in Section 16.1.1. To avoid problems with negative values, we consider a lazy version of the L-reversal chain: at each step, with probability 1/2, perform a uniformly chosen L-reversal, and with probability 1/2, do nothing.

Again, our examplar chain for comparison will be the random transposition chain.

To bound the relaxation time of the L-reversal chain, we must expand each transposition $(a\,b) \in \mathcal{S}_n$ as a product of L-reversals. To show the effect of choice of paths, we try three different strategies and compare the resulting congestion ratios.

We can normalize our presentation of the transposition $(a\,b)$ so that the distance around the cycle from a to b in the positive direction is at most $n/2$. Call the transposition $(a\,b)$ ***short*** when $b = a + k$ for some $k < L$ (interpreted mod n if necessary); call a transposition ***long*** if it is not short. When $b = a + 1$, we have $(a\,b) = \rho_{a,b}$. When $a + 2 \le b \le a + k$, we have $(a\,b) = \rho_{a+1,b-1}\,\rho_{a,b}$. We use these paths of length 1 or 2 for all short transpositions. We will express our other paths below in terms of short transpositions; to complete the expansion, we replace each short transposition with two L-reversals.

Paths for long transpositions, first method. Let $(a\,b)$ be a long transposition. We build $(a\,b)$ by taking the marker at position a on maximal length leaps for as long as we can, then finishing with a correctly-sized jump to get to position b; then take the marker that was at position b over to position a with maximal length leaps. More precisely, write

$$b = a + d(L - 1) + r,$$

with $0 \le r < L - 1$, and set $c_i = a + i(L - 1)$ for $1 \le i \le d$. Then

$$(a\,b) = [(a\ c_1)(c_1\ c_2)\ldots(c_{d-1}\ c_d)]\,(b\ c_d)\,[(c_d\ c_{d-1})\ldots(c_2\ c_1)(c_1\ a)]\,.$$

See Figure 16.4.

Consider the congestion ratio

$$B = \max_{s \in S} \frac{1}{\mu(s)} \sum_{\tilde{s} \in \tilde{S}} \tilde{\mu}(\tilde{s}) N(s, \tilde{s}) \, |\tilde{s}| \leq \max_{\rho_{i,j} \in S} \frac{4L}{n} \sum_{(a\,b) \in \tilde{S}} O\left(\frac{n}{L}\right)$$

of Corollary 13.27. Here S and μ come from the L-reversal walk, while $\tilde{S}$ and $\tilde{\mu}$ come from the random transpositions walk. The initial estimate goes through because the length of all generator paths is at most $O(n/L)$, while any single L-reversal can be used at most twice in a single generator path.

We must still bound the number of different paths in which a particular reversal might appear. This will clearly be maximized for the reversals of length $L-1$, which are used in both the "leaps" of length $L-1$ and the final positioning jumps. Given a reversal $\rho = \rho_{i,i+L-1}$, there are at most $(n/2)/(L-1)$ possible positions for the left endpoint a of a long transposition whose path includes ρ. For each possible left endpoint, there are fewer than $n/2$ possible positions for the right endpoint b (we could bound this more sharply, but it would only save us a factor of 2 to do so). The reversal ρ is also used for short transpositions, but the number of those is only $O(1)$. Hence for this collection of paths we have

$$B = O\left(\frac{n^2}{L}\right).$$

Paths for long transpositions, second method. We now use a similar strategy for moving markers long distances, but try to balance the usage of short transpositions of all available sizes. Write

$$b = a + c\left(\frac{L(L-1)}{2}\right) + r,$$

with $0 \leq r < L(L-1)/2$.

To move the marker at position a to position b, do the following c times: apply the transpositions that move the marker by $L-1$ positions, then by $L-2$ positions, and so on, down to moving 1 position. To cover the last r steps, apply transpositions of lengths $L-1, L-2, \ldots$ until the next in sequence hits exactly or would overshoot; if necessary, apply one more transposition to complete moving the marker to position b. Reverse all but the last transposition to move the marker from position b to position a.

Estimating the congestion ratio works very similarly to the first method. The main difference arises in estimating the number of transpositions $(a\,b)$ whose paths use a particular reversal $\rho = \rho_{i,j}$. Now the left endpoint a can fall at one of at most $2\left(\frac{n/2}{L(L-1)/2}\right)$ positions (the factor of 2 comes from the possibility that ρ is the final jump), since there are at most this number of possible positions for a transposition of the same length as ρ in one of our paths. The right endpoint b again has at most $n/2$ possible values (again, an overestimate that only affects the lead constant). We get

$$B = O\left(\frac{n^2}{L^2}\right). \tag{16.9}$$

That is, we have asymptotically reduced the congestion ratio by a factor of L by changing the paths to use reversals of all sizes evenly.

Paths for long transpositions, third method: randomized. We can use the method described in Remark 13.28 of choosing random, rather than canonical,

paths to match the bound of (16.9). We again describe the paths in terms of short transpositions; to complete the expansion, replace each short transposition with two short reversals.

Fix a transposition $(b\,c)$. Take b on jumps towards c of size uniformly chosen between $L/2+1$ and $L-1$ until it is within distance $L-1$ of c; then make the last jump the required size. To take c back, use the same sequence of jumps, but in reverse.

We must estimate the congestion ratio of (13.28):

$$B = \max_{s\in S} \frac{1}{\mu(s)} \sum_{a\in\tilde{S}} \tilde{\mu}(a) \sum_{\Gamma\in\mathcal{P}_a} \nu_a(\Gamma) N(s,\Gamma)\,|\Gamma|.$$

Since all but the last step of our paths are reversals of length at least $L/2$, for all Γ of positive measure we have $n/L + O(1) < |\Gamma| < 2n/L + O(1)$. Any single reversal can appear at most twice in a single path. Hence

$$B \le \max_{s\in S} \frac{2nL}{n^2} \left(\frac{2n}{L} + O(1)\right) \sum_{a\in\tilde{S}} \sum_{\Gamma\in\mathcal{P}_a} \nu_a(\Gamma).$$

For any $a \in \tilde{S}$, the number of pairs (b,c) for which a can be used is certainly at most n^2. Once we fix b and c, the probability of hitting exactly the span of a while choosing the random path is at most $2(2/L)^2$. (Why? The reversal a is used by at most 2 short transpositions. The probability of choosing the correct left endpoint for one of those transpositions is at most $(2/L)$ (to make this clearer, consider conditioning on all possible partial paths long enough that the left endpoint could possibly be hit). Once the correct left endpoint is hit, the probability of hitting the correct right endpoint is bounded by $2/L$.) Hence for this construction of random paths, we have $B = O(n^2/L^2)$.

REMARK 16.3. Notice that when $L = 2$, all three methods reduce to the paths used in Section 16.1.1 for random adjacent transpositions.

To finish bounding the mixing time, we follow the method of our low-quality estimate (16.1) of $O(n\log n)$ for the relaxation time of the random transposition chain. By Corollary 13.27 and the laziness of the L-reversal chain, we have

$$t_{\text{rel}} = O\left(\frac{n^3 \log n}{L^2}\right)$$

for the L-reversal chain. Finally, as in Section 16.1.1, we use Theorem 12.3 to bound the mixing time by the relaxation time, obtaining

$$t_{\text{mix}} \le \log(4n!)t_{\text{rel}} = O\left(\frac{n^4 \log^2 n}{L^2}\right).$$

Exercise

EXERCISE 16.1. Modify the argument of Proposition 16.2 to cover the case $n/2 < L < n-1$. (Hint: there are now pairs of edges both of which can be broken by two different allowed reversals.)

Notes

Random adjacent transpositions are among the examples analyzed by Diaconis and Saloff-Coste (1993b), who introduced the comparison method for groups. While our presentation uses the same paths and gets the same inequality between the underlying Dirichlet forms, our final bound on the mixing time is much weaker because we apply this inequality only to the spectral gap. Diaconis and Shahshahani (1981) derived very precise information on the spectrum and convergence behavior of the random transpositions walk, and Diaconis and Saloff-Coste (1993b) exploited this data to obtain an $O(n^3 \log n)$ upper bound on the mixing time of the random adjacent transpositions chain.

Diaconis and Saloff-Coste (1993b) proved the first lower bound we present for this chain and conjectured that the upper bound is of the correct asymptotic order. That it is was shown in Wilson (2004a).

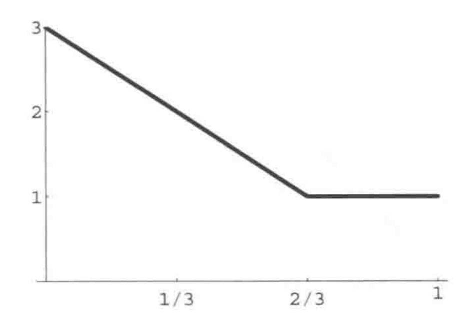


FIGURE 16.5. When $L = n^\alpha$ and $0 < \alpha < 1$, the mixing of the L-reversal chain takes at least $\Omega(n^{1 \vee (3-3\alpha)} \log n)$ steps. This plot shows $1 \vee (3 - 3\alpha)$.

Durrett (2003) introduced the L-reversal chain and proved both bounds we present. For the upper bound, our presentation has again significantly weakened the result by considering only the spectral gap; Durrett proved an upper bound of order $O\left(\frac{n^3 \log n}{L^2}\right)$.

Durrett (2003) also used Wilson's method to give another lower bound, of order $\Omega\left(\frac{n^3 \log n}{L^3}\right)$, when $L \sim n^\alpha$ for some $0 < \alpha < 1$. Taking the maximum of the two lower bounds for L in this range tells us that the mixing of the L-reversal chain takes at least $\Omega(n^{1 \vee (3-3\alpha)} \log n)$ steps—see Figure 16.5. Durrett conjectured that this lower bound is, in fact, sharp.

Cancrini, Caputo, and Martinelli (2006) showed that the relaxation time of the L-reversal chain is $\Theta(n^{1 \vee (3-3\alpha)})$. Morris (2008) has proved an upper bound on the mixing time that is only $O(\log^2 n)$ larger than Durrett's conjecture.

Kandel, Matias, Unger, and Winkler (1996) discuss shuffles relevant to a different problem in genomic sequence analysis.



CHAPTER 17

Martingales and Evolving Sets

17.1. Definition and Examples

Let $X_1, X_2, \ldots$ be i.i.d. $\mathbb{Z}$-valued random variables with $\mathbf{E}(X_t) = 0$ for all t, and let $S_t = \sum_{r=1}^{t} X_r$. The sequence (S_t) is a random walk on $\mathbb{Z}$ with increments (X_t). Given any sequence $(x_1, x_2, \ldots, x_t)$ of possible values for the increments with $\mathbf{P}\{X_1 = x_1, \ldots, X_t = x_t\} > 0$, let $s_t = \sum_{r=1}^{t} x_r$, and observe that the conditional expected position of the walk at time $t+1$ equals s_t, the position at time t:

$$\begin{aligned}\mathbf{E}(S_{t+1} \mid X_1 = x_1, \ldots, X_t = x_t) &= \mathbf{E}(S_t + X_{t+1} \mid X_1 = x_1, \ldots, X_t = x_t) \\ &= \mathbf{E}(X_{t+1} \mid X_1 = x_1, \ldots, X_t = x_t) + \sum_{r=1}^{t} x_r \\ &= s_t. \qquad (17.1)\end{aligned}$$

The equality (17.1) is the key property shared by martingales, defined below.

Often a random sequence is defined using a source of randomness richer than the sequence itself. For example, a step of the lazy random walk on the hypercube can be generated by selecting a coordinate at random and updating the chosen coordinate with a random unbiased bit. The first time all coordinates have been selected at least once is a strong stationary time for the chain (see Example 6.3 and Section 6.5.2). However, this time is a function of the coordinates selected for updates and is *not* a function of the chain. This example illustrates the usefulness of defining a Markov chain on a probability space containing "extra randomness".

In what follows, the sequence of random variables $(Y_t)_{t=0}^{\infty}$ serves as a basic source of randomness. For example, (Y_t) could be an i.i.d. sequence of $\{-1, +1\}$-valued random variables, or a Markov chain. We suppose that each Y_t is a discrete random variable, but make no other assumption about the distribution of this sequence.

A random sequence (X_t) is ***adapted*** to another random sequence (Y_t) if, for each t, there exists a function g_t such that $X_t = g_t(Y_0, \ldots, Y_t)$.

A ***martingale with respect to*** (Y_t) is a sequence of random variables (M_t) satisfying the following conditions:

(i) $\mathbf{E}|M_t| < \infty$ for all t.
(ii) (M_t) is adapted to (Y_t).
(iii) Suppose that $M_t = g_t(Y_0, \ldots, Y_t)$. For all possible values $y_0, \ldots, y_t$ satisfying $\mathbf{P}\{Y_0 = y_0, \ldots, Y_t = y_t\} > 0$, if $m_t = g_t(y_0, \ldots, y_t)$, then

$$\mathbf{E}(M_{t+1} \mid Y_0 = y_0, \ldots, Y_t = y_t) = m_t.$$

Condition (ii) says that M_t is determined by $(Y_0, \ldots, Y_t)$, and condition (iii) says that given the data $(Y_0, \ldots, Y_t)$, the best predictor of M_{t+1} is M_t.

EXAMPLE 17.1. The unbiased random walk (S_t) defined above is a martingale with respect to the increment sequence (X_t). Since $S_t = \sum_{r=1}^{t} X_r$, clearly condition (ii) holds. Condition (iii) is verified in (17.1).

A ***supermartingale*** (M_t) satisfies conditions (i) and (ii) in the definition of a martingale, but instead of (iii), it obeys the inequality

$$\mathbf{E}(M_{t+1} \mid Y_0, \ldots, Y_t) \leq M_t. \tag{17.2}$$

A ***submartingale*** (M_t) satisfies (i) and (ii) and

$$\mathbf{E}(M_{t+1} \mid Y_0, \ldots, Y_t) \geq M_t. \tag{17.3}$$

For a random walk (S_t), the increments $\Delta S_t := S_{t+1} - S_t$ form an independent sequence with $\mathbf{E}(\Delta S_t) = 0$. For a general martingale, the increments also have mean zero, and although not necessarily independent, they are uncorrelated: for $s < t$,

$$\begin{aligned} \mathbf{E}(\Delta M_t \Delta M_s) &= \mathbf{E}\left(\mathbf{E}\left(\Delta M_t \Delta M_s \mid Y_0, Y_1, \ldots, Y_t\right)\right) \\ &= \mathbf{E}\left(\Delta M_s \mathbf{E}\left(\Delta M_t \mid Y_0, Y_1, \ldots, Y_t\right)\right) \\ &= 0. \end{aligned} \tag{17.4}$$

We have used here the fact, immediate from condition (iii) in the definition of a martingale, that

$$\mathbf{E}(\Delta M_t \mid Y_0, \ldots, Y_t) = 0, \tag{17.5}$$

which is stronger than the statement that $\mathbf{E}(\Delta M_t) = 0$.

A useful property of martingales is that

$$\mathbf{E}(M_t) = \mathbf{E}(M_0) \quad \text{for all } t \geq 0.$$

EXAMPLE 17.2. Let (Y_t) be a random walk on $\mathbb{Z}$ which moves up one unit with probability p and down one unit with probability $q := 1 - p$, where $p \neq 1/2$. In other words, given $Y_0, \ldots, Y_t$,

$$\Delta Y_t := Y_{t+1} - Y_t = \begin{cases} 1 & \text{with probability } p, \\ -1 & \text{with probability } q. \end{cases}$$

If $M_t := (q/p)^{Y_t}$, then (M_t) is a martingale with respect to (Y_t). Condition (ii) is clearly satisfied, and

$$\begin{aligned} \mathbf{E}\left((q/p)^{Y_{t+1}} \,\middle|\, Y_0 = y_0, \ldots, Y_t = y_t\right) &= \mathbf{E}\left((q/p)^{y_t}(q/p)^{Y_{t+1}-Y_t} \,\middle|\, Y_0 = y_0, \ldots, Y_t = y_t\right) \\ &= (q/p)^{y_t}\left[p(q/p) + q(q/p)^{-1}\right] \\ &= (q/p)^{y_t}. \end{aligned}$$

EXAMPLE 17.3. Let (Y_t) be as in the previous example, let $\mu := p - q$, and define $M_t := Y_t - \mu t$. The sequence (M_t) is a martingale:

$$\mathbf{E}(M_{t+1} - M_t \mid Y_0, \ldots, Y_t) = p - q - \mu = 0.$$

17.2. Optional Stopping Theorem

A sequence of random variables (A_t) is called ***previsible*** with respect to another sequence of random variables (Y_t) if, for each t, there is a function f_t such that $A_t = f_t(Y_0, \ldots, Y_{t-1})$. The random variable A_t is determined by what has occurred strictly before time t.

Suppose that (M_t) is a martingale with respect to (Y_t) and (A_t) is a previsible sequence with respect to (Y_t). Imagine that a gambler can bet on a sequence of games so that he receives $M_t - M_{t-1}$ for each unit bet on the t-th game. The interpretation of the martingale property $\mathbf{E}(M_t - M_{t-1} \mid Y_0, \ldots, Y_{t-1}) = 0$ is that the games are fair. Let A_t be the amount bet on the t-th game; the fact that the player sizes his bet based only on the outcomes of previous games forces (A_t) to be a previsible sequence. At time t, the gambler's fortune is

$$F_t = M_0 + \sum_{s=1}^{t} A_s(M_s - M_{s-1}). \tag{17.6}$$

Is it possible, by a suitably clever choice of bets $(A_1, A_2, \ldots)$, to generate an advantage for the player? By this, we mean is it possible that $\mathbf{E}(F_t) > 0$ for some t? Many gamblers believe so. Unfortunately, the next theorem proves that they are wrong.

Define for a martingale (M_t) and a previsible sequence (A_t), the sequence of random variables

$$N_t := M_0 + \sum_{s=1}^{t} A_s(M_s - M_{s-1}), \tag{17.7}$$

which is adapted to (Y_t).

THEOREM 17.4. *For any previsible sequence (A_t) such that each A_t is bounded, if (M_t) is a martingale (submartingale) with respect to (Y_t), then the sequence of random variables (N_t) defined in* (17.7) *is also a martingale (submartingale) with respect to (Y_t).*

PROOF. We consider the case where (M_t) is a martingale; the proof when (M_t) is a submartingale is similar.

For each t there is a constant K_t such that $|A_t| \le K_t$, whence

$$\mathbf{E}|N_t| \le \mathbf{E}|M_0| + \sum_{s=1}^{t} K_t \mathbf{E}|M_s - M_{s-1}| < \infty,$$

and therefore the expectation of N_t is defined. Observe that

$$\mathbf{E}\left(N_{t+1} - N_t \mid Y_0, \ldots, Y_t\right) = \mathbf{E}(A_{t+1}(M_{t+1} - M_t) \mid Y_0, \ldots, Y_t).$$

Since A_{t+1} is a function of $Y_0, \ldots, Y_t$, the right-hand side equals

$$A_{t+1}\mathbf{E}(M_{t+1} - M_t \mid Y_0, \ldots, Y_t) = 0.$$

■

Recall from Section 6.2.1 that a stopping time for (Y_t) is a random variable τ with values in $\{0, 1, \ldots\} \cup \{\infty\}$ such that the event $\{\tau = t\}$ is determined by the random variables $Y_0, \ldots, Y_t$. In other words, the sequence of indicator variables $(\mathbf{1}_{\{\tau = t\}})_t$ is adapted to the sequence (Y_t).

For a martingale, $\mathbf{E}(M_t) = \mathbf{E}(M_0)$ for all *fixed* times t. Does this remain valid if we replace t by a random time? In particular, for stopping times τ, is $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$? Under some additional conditions, the answer is "yes". However, as the next example shows, this does not hold in general.

EXAMPLE 17.5. Let (X_s) be the i.i.d. sequence with

$$\mathbf{P}\{X_1 = +1\} = \mathbf{P}\{X_1 = -1\} = \frac{1}{2}.$$

As discussed in Example 17.1, the sequence of partial sums (S_t) is a martingale. We suppose that $S_0 = 0$. The first-passage time to 1, defined as $\tau := \min\{t \geq 0 : S_t = 1\}$, is a stopping time, and clearly $\mathbf{E}(M_\tau) = 1 \neq \mathbf{E}(M_0)$.

Note that if τ is a stopping time, then so is $\tau \wedge t$ for any fixed t.

THEOREM 17.6 (Optional Stopping Theorem, Version 1). *If (M_t) is a martingale and τ is a stopping time, then $(M_{t\wedge\tau})$ is a martingale. Consequently, $\mathbf{E}(M_{t\wedge\tau}) = \mathbf{E}(M_0)$.*

COROLLARY 17.7 (Optional Stopping Theorem, Version 2). *Let (M_t) be a martingale and τ a stopping time. If $\mathbf{P}\{\tau < \infty\} = 1$ and $|M_{t\wedge\tau}| \leq K$ for all t and some constant K, then $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$.*

PROOF OF THEOREM 17.6. If $A_t := \mathbf{1}_{\{\tau \geq t\}}$, then

$$A_t = 1 - \mathbf{1}_{\{\tau \leq t-1\}} = 1 - \sum_{s=1}^{t-1} \mathbf{1}_{\{\tau = s\}}.$$

Since τ is a stopping time, the above equality shows that A_t can be written as a function of $Y_0, \ldots, Y_{t-1}$, whence (A_t) is previsible. By Theorem 17.4,

$$\sum_{s=1}^{t} A_s(M_s - M_{s-1}) = M_{t\wedge\tau} - M_0$$

defines a martingale. The reader should verify that adding M_0 does not destroy the martingale properties, whence $(M_{t\wedge\tau})$ is also a martingale. ■

PROOF OF COROLLARY 17.7. Since $(M_{\tau\wedge t})$ is a martingale, $\mathbf{E}\left(M_{\tau\wedge t}\right) = \mathbf{E}\left(M_0\right)$. Thus

$$\lim_{t\to\infty} \mathbf{E}(M_{\tau\wedge t}) = \mathbf{E}(M_0).$$

By the Bounded Convergence Theorem, the limit and expectation can be exchanged. Since $\mathbf{P}\{\tau < \infty\} = 1$, we have $\lim_{t\to\infty} M_{\tau\wedge t} = M_\tau$ with probability one, and consequently $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$. ■

COROLLARY 17.8 (Optional Stopping Theorem, Version 3). *Let (M_t) be a martingale with bounded increments, that is $|M_{t+1} - M_t| \leq B$ for all t, where B is a non-random constant. Suppose that τ is a stopping time with $\mathbf{E}(\tau) < \infty$. Then $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$.*

PROOF. Note that

$$|M_{\tau\wedge t}| = \left|\sum_{s=1}^{\tau\wedge t}(M_s - M_{s-1}) + M_0\right| \leq \sum_{s=1}^{\tau\wedge t}|M_s - M_{s-1}| + |M_0| \leq B\tau + |M_0|.$$

Since $\mathbf{E}(B\tau + |M_0|) < \infty$, by the Dominated Convergence Theorem and Theorem 17.6,

$$\mathbf{E}(M_0) = \lim_{t\to\infty} \mathbf{E}(M_{\tau\wedge t}) = \mathbf{E}(M_\tau).$$

■

EXAMPLE 17.9. Consider the same set-up as in Example 17.5, so that the partial sums (S_t) associated with i.i.d. unbiased ± 1's is a martingale. Consider the previsible sequence defined by

$$A_t = \begin{cases} 2^t & \text{if } Y_1 = Y_2 = \cdots = Y_{t-1} = -1, \\ 0 & \text{if } Y_s = 1 \text{ for some } s < t. \end{cases}$$

Viewing this sequence as wagers on i.i.d. fair games which pay ± 1 per unit bet, provided the player has not won a single previous game prior to the t-th game, he bets 2^t. At his first win, he stops playing. If τ is the time of the first win, τ is a stopping time. The amount won at time t is

$$M_t := \sum_{s=1}^{t} A_s(M_s - M_{s-1}) = \begin{cases} 0 & \text{if } t = 0, \\ -2^{(t-1)} & \text{if } 1 \le t < \tau, \\ 1 & \text{if } t \ge \tau. \end{cases}$$

Since we are assured that $Y_s = 1$ for some s eventually, $\tau < \infty$ and $M_\tau = 1$. Thus $\mathbf{E}(M_\tau) = 1$. But $\mathbf{E}(M_0) = 0$, and (M_t) is a martingale! By doubling our bets every time we lose, we have assured ourselves of a profit. This at first glance seems to contradict Corollary 17.7. But notice that the condition $|M_{\tau\wedge t}| < K$ is not satisfied, so we cannot apply the corollary.

17.3. Applications

17.3.1. Gambler's ruin. Let (S_t) be a random walk on $\mathbb{Z}$ having ± 1 increments. Define for each integer r the stopping time $\tau_r = \inf\{t \ge 0 : S_t = r\}$, the first time the walk visits r. For $k = 0, 1, \ldots, N$, let

$$\alpha(k) := \mathbf{P}_k\{\tau_0 < \tau_N\}$$

be the probability that the walker started from k visits 0 before hitting N. If a gambler is betting a unit amount on a sequence of games and starts with k units, $\alpha(k)$ is the probability that he goes bankrupt before he attains a fortune of N units.

We suppose that $P\{S_{t+1} - S_t = +1 \mid S_0, \ldots, S_t\} = p$, where $p \neq 1/2$. We use martingales to derive the gambler's ruin formula, which was found previously in Example 9.9 by calculating effective resistance.

In Example 17.2 it was shown that $M_t := (q/p)^{S_t}$ defines a martingale, where $q = 1 - p$. Let $\tau := \tau_0 \wedge \tau_N$ be the first time the walk hits either 0 or N; the random variable τ is a stopping time. Since $M_{\tau\wedge t}$ is bounded, we can apply Corollary 17.7 to get

$$\mathbf{E}_k\left((q/p)^{S_\tau}\right) = (q/p)^k.$$

We can break up the expectation above to get

$$\mathbf{E}_k\left((q/p)^{S_\tau}\right) = \alpha(k) + (q/p)^N(1 - \alpha(k)).$$

Combining these two equations and solving for $\alpha(k)$ yields

$$\alpha(k) = \frac{(q/p)^k - (q/p)^N}{1 - (q/p)^N}.$$

In the case where $p = q = 1/2$, we can apply the same argument to the martingale (S_t) to show that $\alpha(k) = 1 - k/N$.

Now consider again the unbiased random walk. The expected time-to-ruin formula (2.3), which was derived in Section 2.1 by solving a system of linear equations, can also be found using a martingale argument.

Notice that

$$\begin{aligned}\mathbf{E}(S_{t+1}^2 - S_t^2 \mid S_0, \ldots, S_t) &= (S_t + 1)^2 \frac{1}{2} + (S_t - 1)^2 \frac{1}{2} - S_t^2 \\ &= 1,\end{aligned}$$

whence $M_t := S_t^2 - t$ defines a martingale. By the Optional Stopping Theorem (Theorem 17.6),

$$k^2 = \mathbf{E}_k(M_0) = \mathbf{E}_k(M_{\tau \wedge t}) = \mathbf{E}_k(S_{\tau \wedge t}^2) - \mathbf{E}_k(\tau \wedge t). \tag{17.8}$$

Since $(\tau \wedge t) \uparrow \tau$ as $t \to \infty$, the Monotone Convergence Theorem implies that

$$\lim_{t \to \infty} \mathbf{E}_k(\tau \wedge t) = \mathbf{E}_k(\tau). \tag{17.9}$$

Observe that $S_{\tau \wedge t}^2$ is bounded by N^2, so together (17.8) and (17.9) show that

$$\mathbf{E}_k(\tau) = \lim_{t \to \infty} \mathbf{E}_k\left(S_{\tau \wedge t}^2\right) - k^2 \leq N^2 < \infty. \tag{17.10}$$

In particular, this establishes that $\mathbf{P}_k\{\tau < \infty\} = 1$. Therefore, with probability one, $\lim_{t \to \infty} S_{\tau \wedge t}^2 = S_\tau^2$, so by the Dominated Convergence Theorem,

$$\lim_{t \to \infty} \mathbf{E}_k\left(S_{\tau \wedge t}^2\right) = \mathbf{E}_k\left(S_\tau^2\right). \tag{17.11}$$

Taking limits in (17.8) and using (17.9) and (17.11) shows that

$$\mathbf{E}_k \tau = \mathbf{E}_k S_\tau^2 - k^2.$$

Breaking up the expectation $\mathbf{E}_k(S_\tau^2)$ according to whether $\tau = \tau_0$ or $\tau = \tau_N$ yields

$$[1 - \alpha(k)]N^2 - k^2 = \mathbf{E}_k(\tau).$$

Hence,

$$\mathbf{E}_k(\tau) = k(N - k).$$

17.3.2. Waiting times for patterns in coin tossing. Let $X_1, X_2, \ldots$ be a sequence of independent fair coin tosses (so that $\mathbf{P}\{X_t = H\} = \mathbf{P}\{X_t = T\} = 1/2$), and define

$$\tau_{HTH} := \inf\{t \geq 3 : X_{t-2}X_{t-1}X_t = HTH\}.$$

We wish to determine $\mathbf{E}(\tau_{HTH})$.

Gamblers are allowed to place bets on each individual coin toss. On each bet, the gambler is allowed to pay an entrance fee of k units and is payed in return $2k$ units if the outcome is H or 0 units if the outcome is T. The amount k may be negative, which corresponds to a bet on the outcome T.

We suppose that at each unit of time until the word HTH first appears, a new gambler enters and employs the following strategy: on his first bet, he wagers 1 unit on the outcome H. If he loses, he stops. If he wins and the sequence HTH still has not yet appeared, he wagers his payoff of 2 on T. Again, if he loses, he stops playing. As before, if he wins and the sequence HTH has yet to occur, he takes his payoff (now 4) and wagers on H. This is the last bet placed by this particular player.

We describe the situation a bit more precisely: let (B_t) be an i.i.d. sequence of $\{0,1\}$-valued random variables, with $\mathbf{E}(B_t) = 1/2$, and define $M_t = \sum_{s=1}^t (2B_s - 1)$. Clearly (M_t) is a martingale. Let $\tau_{101} = \inf\{t \geq 3 : B_{t-2}B_{t-1}B_t = 101\}$, and define

$$A_t^s = \begin{cases} 1 & t = s, \\ -2 & t = s+1,\ \tau > t, \\ 4 & t = s+2,\ \tau > t, \\ 0 & \text{otherwise.} \end{cases}$$

The random variable $N_t^s = \sum_{r=1}^t A_r^s (M_r - M_{r-1})$ is the profit of the s-th gambler at the t-th game. By Theorem 17.4, the sequence $(N_t^s)_{t=0}^\infty$ is a martingale, and by the Optional Stopping Theorem (Corollary 17.8),

$$\mathbf{E}(N_\tau^s) = 0.$$

Suppose that $\tau_{101} = t$. The gambler who started at t is paid 2 units, the gambler who started at time $t-2$ is paid 8 units, and every gambler has paid an initial 1 entry fee. Since the game is fair, the expected winnings must total 0, so

$$10 - \mathbf{E}(\tau_{101}) = 0.$$

That is, $\mathbf{E}(\tau_{101}) = 10$.

It is (sometimes) surprising to the non-expert that the expected time to see HHH is longer than HTH: modifying the argument above, so that each player bets on the sequence HHH, doubling his bet until he loses, the gambler entering at time $\tau - 2$ is paid 8 units, the gambler entering at time $\tau - 1$ is paid 4 units, and the gambler entering at τ_{HHH} is paid 2. Again, the total outlay is τ_{HHH}, and fairness requires that $\mathbf{E}(\tau_{HHH}) = 8 + 4 + 2 = 14$.

17.4. Evolving Sets

For a *reversible* Markov chain, combining Theorem 12.3 with Theorem 13.14 shows that $t_{\text{mix}}(\varepsilon) \leq -\log(\varepsilon\pi_{\min})t_{\text{rel}}$. Here we give a direct proof for this bound, not requiring reversibility, using evolving sets, a process introduced by Morris and Peres (2005) and defined below.

THEOREM 17.10. *Let P be a lazy irreducible transition matrix, so that $P(x,x) \geq 1/2$ for all $x \in \Omega$, with stationary distribution π. The mixing time $t_{\text{mix}}(\varepsilon)$ satisfies*

$$t_{\text{mix}}(\varepsilon) \leq \frac{2}{\Phi_\star^2} \log\left(\frac{1}{\varepsilon\pi_{\min}}\right).$$

REMARK 17.11. Suppose the chain is reversible. Combining the inequality (17.30), derived in the proof of Theorem 17.10, with the inequality (12.13) yields

$$\frac{|\lambda|^t}{2} \leq d(t) \leq \frac{1}{\pi_{\min}}\left(1 - \frac{\Phi_\star^2}{2}\right)^t,$$

where λ is an eigenvalue of P not equal to 1. Taking the t-th root on the left and right sides above and letting $t \to \infty$ shows that

$$|\lambda| \leq 1 - \frac{\Phi_\star^2}{2},$$

which yields the lower bound in Theorem 13.14 (but restricted to lazy chains).

The proof proceeds by a series of lemmas. Recall that $Q(x,y) = \pi(x)P(x,y)$ and

$$Q(A,B) = \sum_{\substack{x \in A \\ y \in B}} Q(x,y).$$

Observe that $Q(\Omega, y) = \pi(y)$.

The ***evolving-set process*** is a Markov chain on subsets of Ω. Suppose the current state is $S \subset \Omega$. Let U be a random variable which is uniform on $[0,1]$. The next state of the chain is the set

$$\tilde{S} = \left\{ y \in \Omega : \frac{Q(S,y)}{\pi(y)} \geq U \right\}. \tag{17.12}$$

This defines a Markov chain with state space 2^{Ω}, the collection of all subsets of Ω. Note that the chain is not irreducible, because once it hits either the state $\varnothing$ or Ω, it is absorbed. From (17.12), it follows that

$$\mathbf{P}\{y \in S_{t+1} \mid S_t\} = \frac{Q(S_t, y)}{\pi(y)}. \tag{17.13}$$

LEMMA 17.12. *If $(S_t)_{t=0}^{\infty}$ is the evolving-set process associated to the transition matrix P, then*

$$P^t(x,y) = \frac{\pi(y)}{\pi(x)} \mathbf{P}_{\{x\}} \{y \in S_t\}. \tag{17.14}$$

PROOF. We prove this by induction on t. When $t = 0$, both sides of (17.14) equal $\mathbf{1}_{\{y=x\}}$; hence the equality is valid.

Assume that (17.14) holds for $t = s$. By conditioning on the position of the chain after s steps and then using the induction hypothesis, we have that

$$P^{s+1}(x,y) = \sum_{z \in \Omega} P^s(x,z)P(z,y) = \sum_{z \in \Omega} \frac{\pi(z)}{\pi(x)} \mathbf{P}_{\{x\}}\{z \in S_s\} P(z,y). \tag{17.15}$$

By switching summation and expectation,

$$\begin{aligned}\sum_{z \in \Omega} \pi(z) \mathbf{P}_{\{x\}}\{z \in S_s\} P(z,y) &= \sum_{z \in \Omega} \mathbf{E}_{\{x\}} \left(\mathbf{1}_{\{z \in S_s\}} \pi(z) P(z,y) \right) \\ &= \mathbf{E}_{\{x\}} \left(\sum_{z \in S_s} Q(z,y) \right) = \mathbf{E}_{\{x\}} \left(Q(S_s, y) \right). \end{aligned} \tag{17.16}$$

From (17.13), (17.15), and (17.16),

$$P^{s+1}(x,y) = \frac{1}{\pi(x)} \mathbf{E}_{\{x\}} \left(\pi(y) \mathbf{P}\{y \in S_{s+1} \mid S_s\} \right) = \frac{\pi(y)}{\pi(x)} \mathbf{P}_{\{x\}}\{y \in S_{s+1}\}.$$

Thus, (17.14) is proved for $t = s+1$, and by induction, it must hold for all t. ■

LEMMA 17.13. *The sequence $\{\pi(S_t)\}$ is a martingale.*

PROOF. We have

$$\mathbf{E}(\pi(S_{t+1}) \mid S_t) = \mathbf{E}\left(\sum_{z \in \Omega} \mathbf{1}_{\{z \in S_{t+1}\}} \pi(z) \,\middle|\, S_t \right).$$

By (17.13), the right-hand side above equals

$$\sum_{z\in\Omega}\mathbf{P}\{z\in S_{t+1}\mid S_t\}\pi(z)=\sum_{z\in\Omega}Q(S_t,z)=Q(S_t,\Omega)=\pi(S_t),$$

which concludes the proof. ∎

Recall that $\Phi(S)=Q(S,S^c)/\pi(S)$ is the bottleneck ratio of the set S, defined in Section 7.2.

LEMMA 17.14. *Let $R_t=\pi(S_{t+1})/\pi(S_t)$, and let (U_t) be a sequence of independent random variables, each uniform on $[0,1]$, such that S_{t+1} is generated from S_t using U_{t+1}. Recall that $\Phi(S)=Q(S,S^c)/\pi(S)$. Then*

$$\mathbf{E}(R_t\mid U_{t+1}\le 1/2,\ S_t=S)=1+2\Phi(S), \tag{17.17}$$

$$\mathbf{E}(R_t\mid U_{t+1}> 1/2,\ S_t=S)=1-2\Phi(S). \tag{17.18}$$

PROOF. Since the chain is lazy, $Q(y,y)\ge\pi(y)/2$, so if $y\notin S$,

$$\begin{aligned}\frac{Q(S,y)}{\pi(y)}=\sum_{x\in S}\frac{Q(x,y)}{\pi(y)}&\le\sum_{\substack{x\in\Omega\\ x\ne y}}\frac{Q(x,y)}{\pi(y)}\\&=\sum_{x\in\Omega}\frac{Q(x,y)}{\pi(y)}-\frac{Q(y,y)}{\pi(y)}=1-\frac{Q(y,y)}{\pi(y)}\le\frac{1}{2}.\end{aligned} \tag{17.19}$$

Given $U_{t+1}\le 1/2$, the distribution of U_{t+1} is uniform on $[0,1/2]$. By (17.19), for $y\notin S$,

$$\mathbf{P}\left\{\frac{Q(S,y)}{\pi(y)}\ge U_{t+1}\,\middle|\,U_{t+1}\le 1/2,\ S_t=S\right\}=2\frac{Q(S,y)}{\pi(y)}.$$

Since $y\in S_{t+1}$ if and only if $U_{t+1}\le Q(S_t,y)/\pi(y)$,

$$\mathbf{P}\{y\in S_{t+1}\mid U_{t+1}\le 1/2,\ S_t=S\}=\frac{2Q(S,y)}{\pi(y)}\quad\text{for } y\notin S. \tag{17.20}$$

Also, since $Q(S,y)/\pi(y)\ge Q(y,y)/\pi(y)\ge 1/2$ for $y\in S$, it follows that

$$\mathbf{P}\{y\in S_{t+1}\mid U_{t+1}\le 1/2,\ S_t=S\}=1\quad\text{for } y\in S. \tag{17.21}$$

We have

$$\begin{aligned}\mathbf{E}\left(\pi(S_{t+1})\mid U_{t+1}\le 1/2,\ S_t=S\right)&=\mathbf{E}\left(\sum_{y\in\Omega}\mathbf{1}_{\{y\in S_{t+1}\}}\pi(y)\,\middle|\,U_{t+1}\le 1/2, S_t=S\right)\\&=\sum_{y\in S}\pi(y)\mathbf{P}\{y\in S_{t+1}\mid U_{t+1}\le 1/2,\ S_t=S\}\\&\quad+\sum_{y\notin S}\pi(y)\mathbf{P}\{y\in S_{t+1}\mid U_{t+1}\le 1/2,\ S_t=S\}.\end{aligned}$$

By the above, (17.20), and (17.21),

$$\mathbf{E}\left(\pi(S_{t+1})\mid U_{t+1}\le 1/2,\ S_t=S\right)=\pi(S)+2Q(S,S^c). \tag{17.22}$$

By Lemma 17.13 and (17.22),

$$\begin{aligned}\pi(S) &= \mathbf{E}(\pi(S_{t+1}) \mid S_t = S) \\ &= \frac{1}{2}\mathbf{E}(\pi(S_{t+1}) \mid U_{t+1} \le 1/2,\, S_t = S) + \frac{1}{2}\mathbf{E}(\pi(S_{t+1}) \mid U_{t+1} > 1/2,\, S_t = S) \\ &= \frac{\pi(S)}{2} + Q(S, S^c) + \frac{1}{2}\mathbf{E}(\pi(S_{t+1}) \mid U_{t+1} > 1/2,\, S_t = S).\end{aligned}$$

Rearranging shows that

$$\mathbf{E}(\pi(S_{t+1}) \mid U_{t+1} > 1/2,\, S_t = S) = \pi(S) - 2Q(S, S^c). \tag{17.23}$$

Dividing both sides of (17.22) and (17.23) by $\pi(S)$ yields (17.17) and (17.18), respectively. ■

LEMMA 17.15. *For $\alpha \ge 0$,*

$$\frac{\sqrt{1+2\alpha} + \sqrt{1-2\alpha}}{2} \le \sqrt{1-\alpha^2} \le 1 - \frac{\alpha^2}{2}.$$

This is seen by squaring each side of both inequalities.

LEMMA 17.16. *Let (S_t) be the evolving-set process. If*

$$S_t^\sharp = \begin{cases} S_t & \text{if } \pi(S_t) \le 1/2, \\ S_t^c & \text{otherwise,} \end{cases} \tag{17.24}$$

then

$$\mathbf{E}\left(\sqrt{\pi(S_{t+1}^\sharp)/\pi(S_t^\sharp)} \;\middle|\; S_t \right) \le 1 - \frac{\Phi_\star^2}{2}. \tag{17.25}$$

PROOF. First, letting $R_t := \pi(S_{t+1})/\pi(S_t)$, applying Jensen's inequality shows that

$$\begin{aligned}\mathbf{E}\left(\sqrt{R_t} \mid S_t = S\right) &= \frac{\mathbf{E}\left(\sqrt{R_t} \,\middle|\, U_{t+1} \le 1/2,\, S_t = S\right) + \mathbf{E}\left(\sqrt{R_t} \,\middle|\, U_{t+1} > 1/2,\, S_t = S\right)}{2} \\ &\le \frac{\sqrt{\mathbf{E}\left(R_t \mid U_{t+1} \le 1/2,\, S_t = S\right)} + \sqrt{\mathbf{E}\left(R_t \mid U_{t+1} > 1/2,\, S_t = S\right)}}{2}.\end{aligned}$$

Applying Lemma 17.14 and Lemma 17.15 shows that, for $\pi(S) \le 1/2$,

$$\mathbf{E}\left(\sqrt{R_t} \,\middle|\, S_t = S\right) \le \frac{\sqrt{1+2\Phi(S)} + \sqrt{1-2\Phi(S)}}{2} \le 1 - \frac{\Phi(S)^2}{2} \le 1 - \frac{\Phi_\star^2}{2}. \tag{17.26}$$

Now assume that $\pi(S_t) \le 1/2$. Then

$$\sqrt{\pi(S_{t+1}^\sharp)/\pi(S_t^\sharp)} = \sqrt{\pi(S_{t+1}^\sharp)/\pi(S_t)} \le \sqrt{\pi(S_{t+1})/\pi(S_t)},$$

and (17.25) follows from (17.26). If $\pi(S_t) > 1/2$, then replace S_t by S_t^c in the previous argument. (If (S_t) is an evolving-set process started from S, then (S_t^c) is also an evolving-set process started from S^c.) ■

PROOF OF THEOREM 17.10. From Lemma 17.16,

$$\mathbf{E}\left(\sqrt{\pi(S_{t+1}^\sharp)}\right) \le \mathbf{E}\left(\sqrt{\pi(S_t^\sharp)}\right)\left(1 - \frac{\Phi_\star^2}{2}\right).$$

Iterating,

$$\mathbf{E}_S\left(\sqrt{\pi(S_t^\sharp)}\right) \le \left(1 - \frac{\Phi_\star^2}{2}\right)^t \sqrt{\pi(S)}.$$

Since $\sqrt{\pi_{\min}}\,\mathbf{P}_S\{S_t^\sharp \neq \varnothing\} \leq \mathbf{E}_S\left(\sqrt{\pi(S_t^\sharp)}\right)$, we have

$$\mathbf{P}_S\{S_t^\sharp \neq \varnothing\} \leq \sqrt{\frac{\pi(S)}{\pi_{\min}}}\left(1 - \frac{\Phi_\star^2}{2}\right)^t. \tag{17.27}$$

Since $\{S_t^\sharp \neq \varnothing\} \supset \{S_{t+1}^\sharp \neq \varnothing\}$, by (17.27),

$$\mathbf{P}_S\{S_t^\sharp \neq \varnothing \text{ for all } t \geq 0\} = \mathbf{P}_S\left(\bigcap_{t=1}^{\infty}\{S_t^\sharp \neq \varnothing\}\right) = \lim_{t\to\infty}\mathbf{P}_S\{S_t^\sharp \neq \varnothing\} = 0.$$

That is, $(S_t^\sharp)$ is eventually absorbed in the state $\varnothing$. Let

$$\tau = \min\{t \geq 0 \,:\, S_t^\sharp = \varnothing\}.$$

We have $S_\tau \in \{\varnothing, \Omega\}$ and $\pi(S_\tau) = \mathbf{1}_{\{S_\tau = \Omega\}}$. Note that by Lemma 17.13 and the Optional Stopping Theorem (Corollary 17.7),

$$\pi(x) = \mathbf{E}_{\{x\}}(\pi(S_0)) = \mathbf{E}_{\{x\}}(\pi(S_\tau)) = \mathbf{P}_{\{x\}}\{S_\tau = \Omega\}. \tag{17.28}$$

By (17.28) and Lemma 17.12,

$$\begin{aligned}|P^t(x,y) - \pi(y)| &= \frac{\pi(y)}{\pi(x)}\left|\mathbf{P}_{\{x\}}\{y \in S_t\} - \pi(x)\right| \\ &= \frac{\pi(y)}{\pi(x)}\left|\mathbf{P}_{\{x\}}\{y \in S_t\} - \mathbf{P}_{\{x\}}\{S_\tau = \Omega\}\right|. \end{aligned} \tag{17.29}$$

Using the identity

$$\begin{aligned}\mathbf{P}_{\{x\}}\{y \in S_t\} &= \mathbf{P}_{\{x\}}\{y \in S_t,\, \tau > t\} + \mathbf{P}_{\{x\}}\{y \in S_t, \tau \leq t\} \\ &= \mathbf{P}_{\{x\}}\{y \in S_t, \tau > t\} + \mathbf{P}_{\{x\}}\{S_\tau = \Omega, \tau \leq t\}\end{aligned}$$

in (17.29) shows that

$$\begin{aligned}|P^t(x,y) - \pi(y)| &= \frac{\pi(y)}{\pi(x)}\left|\mathbf{P}_{\{x\}}\{y \in S_t,\, \tau > t\} - \mathbf{P}_{\{x\}}\{S_\tau = \Omega,\, \tau > t\}\right| \\ &\leq \frac{\pi(y)}{\pi(x)}\mathbf{P}_{\{x\}}\{\tau > t\}.\end{aligned}$$

Combining with (17.27),

$$d(t) \leq s(t) \leq \max_{x,y}\frac{|P^t(x,y) - \pi(y)|}{\pi(y)} \leq \frac{1}{\pi_{\min}}\left(1 - \frac{\Phi_\star^2}{2}\right)^t. \tag{17.30}$$

It follows that if $t \geq \frac{2}{\Phi_\star^2}\log\left(\frac{1}{\varepsilon\pi_{\min}}\right)$, then $d(t) \leq \varepsilon$. ■

17.5. A General Bound on Return Probabilities

The goal in this section is to prove the following:

THEOREM 17.17. *Let P be the transition matrix for a lazy random walk on a graph of maximal degree Δ. Then*

$$|P^t(x,x) - \pi(x)| \leq \frac{\sqrt{2}\Delta^{5/2}}{\sqrt{t}}. \tag{17.31}$$

REMARK 17.18. The dependence on Δ in (17.31) is not the best possible. It can be shown that an upper bound of $c_1\Delta/\sqrt{t}$ holds.

We will need the following result about martingales, which is itself of independent interest:

PROPOSITION 17.19. *Let M_t be a non-negative martingale with respect to (Y_t), and define*

$$T_h := \min\{t \geq 0 : M_t = 0 \text{ or } M_t \geq h\}.$$

Assume that

(i) $\text{Var}(M_{t+1} \mid Y_0, \ldots, Y_t) \geq \sigma^2$, *and*
(ii) $M_{T_h} \leq Dh$.

Let $T := T_1$. If M_0 is a constant, then

$$\mathbf{P}\{T > t\} \leq \frac{2M_0}{\sigma}\sqrt{\frac{D}{t}}. \tag{17.32}$$

PROOF. We have that $\{T \geq t\} \subseteq \{T_h \geq t\} \cup \{M_{T_h} \geq h\}$, whence

$$\mathbf{P}\{T \geq t\} \leq \mathbf{P}\{T_h \geq t\} + \mathbf{P}\{M_{T_h} \geq h\}. \tag{17.33}$$

We first bound $\mathbf{P}\{M_{T_h} \geq h\}$. Since $(M_{t \wedge T_h})$ is bounded, by the Optional Stopping Theorem,

$$M_0 \geq \mathbf{E} M_{T_h} \geq h\mathbf{P}\{M_{T_h} \geq h\},$$

whence

$$\mathbf{P}\{M_{T_h} \geq h\} \leq \frac{M_0}{h}. \tag{17.34}$$

We now bound $\mathbf{P}\{T_h > t\}$. Let $G_t := M_t^2 - hM_t - \sigma^2 t$. The sequence (G_t) is a submartingale. Note that for $t \leq T_h$,

$$M_t^2 - hM_t = (M_t - h)M_t \leq (D-1)hM_t;$$

therefore,

$$\mathbf{E}(M_{t\wedge T_h}^2 - hM_{t\wedge T_h}) \leq (D-1)hM_0.$$

Since $(G_{t\wedge T_h})$ is a submartingale,

$$\begin{aligned} -hM_0 \leq G_0 \leq \mathbf{E} G_{t\wedge T_h} &= \mathbf{E}(M_{t\wedge T_h}^2 - hM_{t\wedge T_h}) - \sigma^2 \mathbf{E}(t \wedge T_h) \\ &\leq (D-1)hM_0 - \sigma^2\mathbf{E}(t\wedge T_h). \end{aligned}$$

We conclude that $\mathbf{E}(t \wedge T_h) \leq \frac{DhM_0}{\sigma^2}$. Letting $t \to \infty$, by the Monotone Convergence Theorem, $\mathbf{E}T_h \leq \frac{DhM_0}{\sigma^2}$. By Markov's inequality,

$$\mathbf{P}\{T_h \geq t\} \leq \frac{DhM_0}{\sigma^2 t}.$$

Combining the above bound with with (17.33) and (17.34) shows that

$$\mathbf{P}\{T > t\} \leq \frac{M_0}{h} + \frac{DhM_0}{\sigma^2 t}.$$

We take $h = \sqrt{t\sigma^2/D}$ to optimize the above bound. This proves the inequality (17.32). ■

Many variants of the above proposition are useful in applications. We state one here.

PROPOSITION 17.20. *Let $(Z_t)_{t\geq 0}$ be a non-negative supermartingale, adapted to the sequence (Y_t), and let τ be a stopping time for the sequence (Y_t). Define the random vector $\mathbf{Y}_t := (Y_0, \ldots, Y_t)$. Suppose that*

(i) $Z_0 = k$,
(ii) $\mathbf{E}(Z_{t+1} - Z_t \mid \boldsymbol{Y}_t) \leq B$,
(iii) *there exists a constant* $\sigma^2 > 0$ *such that* $\mathrm{Var}(Z_{t+1} \mid Y_0, Y_1, \ldots, Y_t) > \sigma^2$ *on the event* $\{\tau > t\}$.

If $u > 4B^2/(3\sigma^2)$, *then*

$$\mathbf{P}_k\{\tau > u\} \leq \frac{4k}{\sigma\sqrt{u}}.$$

The proof follows the same outline as the proof of Proposition 17.19 and is left to the reader in Exercise 17.3

We now prove the principal result of this section.

Proof of Theorem 17.17. Let (S_t) be the evolving-set process associated to the Markov chain with transition matrix P. Define

$$\tau := \min\{t \geq 0 : S_t \in \{\varnothing, \Omega\}\}.$$

Observe that, since $\pi(S_t)$ is a martingale,

$$\pi(x) = \mathbf{E}_{\{x\}}\pi(S_0) = \mathbf{E}_{\{x\}}\pi(S_\tau) = \mathbf{E}_{\{x\}}\pi(S_\tau)\mathbf{1}_{\{S_\tau=\Omega\}} = \mathbf{P}_{\{x\}}\{x \in S_\tau\}.$$

By Lemma 17.12, $P^t(x,x) = \mathbf{P}_{\{x\}}\{x \in S_t\}$. Therefore,

$$|P^t(x,x) - \pi(x)| = |\mathbf{P}_{\{x\}}\{x \in S_t\} - \mathbf{P}_{\{x\}}\{x \in S_\tau\}| \leq \mathbf{P}_{\{x\}}\{\tau > t\}.$$

Since conditioning always reduces variance,

$$\mathrm{Var}_S(\pi(S_1)) \geq \mathrm{Var}_S\left(\mathbf{E}(\pi(S_1) \mid \mathbf{1}_{\{U_1 \leq 1/2\}})\right).$$

Note that (see Lemma 17.14)

$$\mathbf{E}_S(\pi(S_1) \mid \mathbf{1}_{\{U_1 \leq 1/2\}}) = \begin{cases} \pi(S) + 2Q(S,S^c) & \text{with probability } 1/2, \\ \pi(S) - 2Q(S,S^c) & \text{with probability } 1/2. \end{cases}$$

Therefore, provided $S \notin \{\varnothing, \Omega\}$,

$$\mathrm{Var}_S\left(\mathbf{E}(\pi(S_1) \mid \mathbf{1}_{\{U_1 \leq 1/2\}})\right) = 4Q(S,S^c)^2 \geq \frac{1}{n^2\Delta^2}.$$

The last inequality follows since $\pi(x) = \deg(x)/(2E)$, where E is the number of edges in the graph, and if $S \notin \{\varnothing, \Omega\}$, then there exists x, y such that $x \in S$, $y \notin S$ and $P(x,y) > 0$, whence

$$Q(S,S^c) \geq \pi(x)P(x,y) \geq \frac{\deg(x)}{2E}\frac{1}{2\deg(x)} \geq \frac{1}{4E} \geq \frac{1}{2n\Delta}.$$

Note that $\pi(S_{t+1}) \leq (\Delta+1)\pi(S_t)$. Therefore, we can apply Proposition 17.19 with $D = \Delta + 1$ and $M_0 \leq \Delta/n$ to obtain the inequality (17.31). ∎

17.6. Harmonic Functions and the Doob h-Transform

Recall that a function $h : \Omega \to \mathbb{R}$ is harmonic for P if $Ph = h$. The connection between harmonic functions, Markov chains, and martingales is that if (X_t) is a Markov chain with transition matrix P and h is a P-harmonic function, then $M_t = h(X_t)$ defines a martingale with respect to (X_t):

$$\begin{aligned}\mathbf{E}\left(M_{t+1} \mid X_0, X_1, \ldots, X_t\right) &= \mathbf{E}\left(M_{t+1} \mid X_t\right) \\ &= \sum_{y \in \Omega} P(X_t, y)h(y) = Ph(X_t) = h(X_t) = M_t.\end{aligned}$$

17.6.1. Conditioned Markov chains and the Doob transform. Let P be a Markov chain such that the set B is absorbing: $P(x,x) = 1$ for $x \in B$. We allow $B = \varnothing$. Let h be a positive harmonic function on $\Omega \setminus B$, and define

$$\check{P}(x,y) := \frac{P(x,y)h(y)}{h(x)}.$$

Note that for $x \notin B$,

$$\sum_{y\in\Omega} \check{P}(x,y) = \frac{1}{h(x)} \sum_{y\in\Omega} h(y)P(x,y) = \frac{Ph(x)}{h(x)} = 1.$$

If $x \in B$, then $\check{P}(x,x) = 1$. Therefore, $\check{P}$ is a transition matrix, called the ***Doob h-transform*** of P.

Let P be a transition matrix, and assume that the states a and b are absorbing. Let $h(x) := \mathbf{P}_x\{\tau_b < \tau_a\}$, and assume that $h(x) > 0$ for $x \neq a$. Since $h(x) = \mathbf{E}_x \mathbf{1}_{\{X_{\tau_a \wedge \tau_b} = b\}}$, Proposition 9.1 shows that h is harmonic on $\Omega \setminus \{a,b\}$, whence we can define the Doob h-transform $\check{P}$ of P. Observe that for $x \neq a$,

$$\begin{aligned}\check{P}(x,y) &= \frac{P(x,y)\mathbf{P}_y\{\tau_b < \tau_a\}}{\mathbf{P}_x\{\tau_b < \tau_a\}} \\ &= \frac{\mathbf{P}_x\{X_1 = y,\ \tau_b < \tau_a\}}{\mathbf{P}_x\{\tau_b < \tau_a\}} \\ &= \mathbf{P}_x\{X_1 = y \mid \tau_b < \tau_a\},\end{aligned}$$

so the chain with matrix $\check{P}$ is the original chain conditioned to hit b before a.

EXAMPLE 17.21 (Conditioning the evolving-set process). Given a transition matrix P on Ω, consider the corresponding evolving-set process (S_t). Let $\tau := \min\{t : S_t \in \{\varnothing, \Omega\}\}$. Since $\{\pi(S_t)\}$ is a martingale, the Optional Stopping Theorem implies that

$$\pi(A) = \mathbf{E}_A \pi(S_\tau) = \mathbf{P}_A\{S_\tau = \Omega\}.$$

If K is the transition matrix of (S_t), then the Doob transform of (S_t) conditioned to be absorbed in Ω has transition matrix

$$\check{K}(A,B) = \frac{\pi(B)}{\pi(A)} K(A,B). \tag{17.35}$$

EXAMPLE 17.22 (Simple random walk on $\{0,1,\ldots,n\}$). Consider the simple random walk on $\{0,1,\ldots,n\}$ with loops on the endpoints:

$$P(0,0) = P(0,1) = \frac{1}{2} \quad \text{and} \quad P(n,n) = P(n,n-1) = \frac{1}{2}.$$

Consider the process conditioned to absorb at n before 0. Since $\mathbf{P}_k\{\tau_n < \tau_0\} = k/n$, we have

$$\check{P}(x,y) = \frac{y}{x} P(x,y) \quad \text{for } 0 < x < n.$$

Note that transition probabilities from 0 and n are irrelevant.

17.7. Strong Stationary Times from Evolving Sets

The goal of this section is to construct a strong stationary time by coupling a Markov chain with the conditioned evolving-set process of Example 17.21.

This construction is due to Diaconis and Fill (1990) and preceded the definition of evolving sets, and thus our notation differs.

The idea is to start with $X_0 = x$ and $S_0 = \{x\}$ and run the Markov chain (X_t) and the evolving-set process (S_t) together, at each stage conditioning on $X_t \in S_t$.

Let P be an irreducible transition matrix, and let K be the transition matrix for the associated evolving-set process. The matrix $\check{K}$ denotes the evolving-set process conditioned to be absorbed in Ω. (See Example 17.21.)

For $y \in \Omega$, define the transition matrix on 2^Ω by

$$J_y(A, B) := \mathbf{P}_A\{S_1 = B \mid y \in S_1\}\mathbf{1}_{\{y \in B\}}.$$

From (17.13) it follows that $J_y(A, B) = K(A, B)\pi(y)\mathbf{1}_{\{y\in B\}}/Q(A, y)$. Define the transition matrix $P^\star$ on $\Omega \times 2^\Omega$ by

$$\begin{aligned} P^\star((x, A), (y, B)) &:= P(x, y)J_y(A, B) \\ &= \frac{P(x, y)K(A, B)\pi(y)\mathbf{1}_{\{y\in B\}}}{Q(A, y)}. \end{aligned}$$

Let (X_t, S_t) be a Markov chain with transition matrix $P^\star$, and let $\mathbf{P}^\star$ denote the probability measure on the space where (X_t, S_t) is defined.

Observe that

$$\sum_{B\,:\,y\in B} P^\star\left((x, A), (y, B)\right) = P(x, y)\frac{\pi(y)}{Q(A, y)} \sum_{B\,:\,y\in B} K(A, B). \tag{17.36}$$

The sum $\sum_{B\,:\,y\in B} K(A, B)$ is the probability that the evolving-set process started from A contains y at the next step. By (17.13) this equals $Q(A, y)/\pi(y)$, whence (17.36) says that

$$\sum_{B\,:\,y\in B} P^\star\left((x, A), (y, B)\right) = P(x, y). \tag{17.37}$$

It follows that (X_t) is a Markov chain with transition matrix P.

THEOREM 17.23 (Diaconis and Fill (1990)). *We abbreviate* $\mathbf{P}^\star_{x,\{x\}}$ *by* $\mathbf{P}^\star_x$.

(i) *If* (X_t, S_t) *is a Markov chain with transition matrix* $P^\star$ *started from* $(x, \{x\})$, *then the sequence* (S_t) *is a Markov chain started from* $\{x\}$ *with transition matrix* $\check{K}$.

(ii) *For* $w \in S_t$,

$$\mathbf{P}^\star_x\{X_t = w \mid S_0, \ldots, S_t\} = \frac{\pi(w)}{\pi(S_t)}.$$

PROOF. We use induction on t. When $t = 0$, both (i) and (ii) are obvious. For the induction step, we assume that for some $t \geq 0$, the sequence $(S_j)_{j=0}^t$ is a Markov chain with transition matrix $\check{K}$ and that (ii) holds. Our goal is to verify (i) and (ii) with $t + 1$ in place of t.

We write $\boldsymbol{S}_t$ for the vector $(S_0, \ldots, S_t)$. For $v \in B$,

$$\begin{aligned}
&\mathbf{P}^\star_x\{X_{t+1} = v,\, S_{t+1} = B \mid \boldsymbol{S}_t\} \\
&\quad = \sum_{w \in S_t} \mathbf{P}^\star\{X_{t+1} = v,\, S_{t+1} = B \mid X_t = w,\, \boldsymbol{S}_t\}\mathbf{P}^\star_x\{X_t = w \mid \boldsymbol{S}_t\}. \quad (17.38)
\end{aligned}$$

Because the process (X_t, S_t) is a Markov chain with transition matrix $P^\star$,

$$\begin{aligned}
\mathbf{P}^\star\{X_{t+1} = v,\, S_{t+1} = B \mid X_t = w,\, \boldsymbol{S}_t\} &= P^\star\left((w, S_t), (v, B)\right) \\
&= \frac{P(w, v)K(S_t, B)\pi(v)}{Q(S_t, v)}. \quad (17.39)
\end{aligned}$$

Substituting the identity (17.39) in the right-hand side of (17.38) and using the induction hypothesis shows that, for $v \in B$,

$$\begin{aligned}
\mathbf{P}^\star_x\{X_{t+1} = v,\, S_{t+1} = B \mid \boldsymbol{S}_t\} &= \sum_{w \in S_t} \frac{P(w, v)K(S_t, B)\pi(v)}{Q(S_t, v)} \frac{\pi(w)}{\pi(S_t)} \\
&= \frac{\pi(v)}{\pi(S_t)} \frac{\sum_{w \in S_t} \pi(w)P(w, v)}{Q(S_t, v)} K(S_t, B) \\
&= \frac{\pi(v)}{\pi(S_t)} K(S_t, B). \quad (17.40)
\end{aligned}$$

Summing over $v \in B$ gives

$$\begin{aligned}
\mathbf{P}^\star_x\{S_{t+1} = B \mid S_0, \ldots, S_t\} &= \frac{\pi(B)K(S_t, B)}{\pi(S_t)} \quad (17.41) \\
&= \check{K}(S_t, B), \quad (17.42)
\end{aligned}$$

where (17.42) follows from (17.35). Therefore, $(S_j)_{j=0}^{t+1}$ is a Markov chain with transition matrix $\check{K}$, which verifies (ii) with $t + 1$ replacing t.

Taking the ratio of (17.40) and (17.41) shows that

$$\mathbf{P}^\star_x\{X_{t+1} = v \mid \boldsymbol{S}_t,\, S_{t+1} = B\} = \frac{\pi(v)}{\pi(B)},$$

which completes the induction step. ■

Corollary 17.24. *For the coupled process (X_t, S_t), consider the absorption time*

$$\tau^\star := \min\{t \geq 0 \,:\, S_t = \Omega\}.$$

Then $\tau^\star$ is a strong stationary time for (X_t).

Proof. This follows from Theorem 17.23(ii): summing over all sequences of sets $(A_1, \ldots, A_t)$ with $A_i \neq \Omega$ for $i < t$ and $A_t = \Omega$,

$$\begin{aligned}
\mathbf{P}^\star_x\{\tau^\star = t,\, X_t = w\} &= \sum \mathbf{P}^\star_x\{(S_1, \ldots, S_t) = (A_1, \ldots, A_t),\, X_t = w\} \\
&= \sum \mathbf{P}^\star_x\{(S_1, \ldots, S_t) = (A_1, \ldots, A_t)\}\pi(w) \\
&= \mathbf{P}^\star\{\tau^\star = t\}\pi(w).
\end{aligned}$$

■

EXAMPLE 17.25. Suppose the base Markov chain is simple random walk on $\{0, 1, \ldots, n\}$ with loops at 0 and n; the stationary distribution π is uniform. In this case we have $S_t = [0, Y_t)$, where (Y_t) satisfies

$$\begin{aligned}\mathbf{P}\{Y_{t+1} = Y_t + 1 \mid Y_t\} &= \mathbf{P}\{Y_t \in S_{t+1} \mid S_t = [0, Y_t)\} \\ &= \frac{1}{2} \\ &= \mathbf{P}\{Y_{t+1} = Y_t - 1 \mid Y_t\}.\end{aligned}$$

Therefore, (Y_t) is a simple random walk on $\{0, \ldots, n+1\}$ with absorption at endpoints.

We deduce that the absorption time $\tau^\star$ when started from $S_0 = \{0\}$ is the absorption time of the simple random walk (Y_t) conditioned to hit $n+1$ before 0 when started at $Y_0 = 1$. Thus

$$\mathbf{E}^\star \tau^\star = \frac{(n+1)^2 - 1}{3} = \frac{n^2 + 2n}{3}.$$

Since, by Corollary 17.24, $\tau^\star$ is a strong stationary time for (X_t), we conclude that $t_{\text{mix}} = O(n^2)$.

Exercises

EXERCISE 17.1. Let (X_t) be the simple random walk on $\mathbb{Z}$.

(a) Show that $M_t = X_t^3 - 3tX_t$ is a martingale.

(b) Let τ be the time when the walker first visits either 0 or n. Show that for $0 \leq k \leq n$,

$$\mathbf{E}_k(\tau \mid X_\tau = n) = \frac{n^2 - k^2}{3}.$$

EXERCISE 17.2. Let (X_t) be a supermartingale. Show that there is a martingale (M_t) and a non-decreasing and previsible sequence (A_t) so that $X_t = M_t - A_t$. This is called the ***Doob decomposition*** of (X_t).

EXERCISE 17.3. Prove Proposition 17.20.

Hint: Use the Doob decomposition $Z_t = M_t - A_t$ (see Exercise 17.2), and modify the proof of Proposition 17.19 applied to M_t.

EXERCISE 17.4. For lazy birth-and-death chains, the evolving-set process started with $S_0 = \{0\}$ always has $S_t = [0, Y_t)$ or $S_t = \varnothing$.

Notes

Doob was the first to call processes that satisfy the conditional expectation property

$$\mathbf{E}(M_{t+1} \mid M_1, \ldots, M_t) = M_t$$

"martingales". The term was used previously by gamblers to describe certain betting schemes.

See Williams (1991) for a friendly introduction to martingales and Doob (1953) for a detailed history.

For much more on the waiting time for patterns in coin tossing, see Li (1980).

Evolving sets. Define $\Phi(r)$ for $r \in [\pi_{\min}, 1/2]$ by

$$\Phi(r) := \inf\{\Phi(S) : \pi(S) \le r\}. \tag{17.43}$$

For reversible, irreducible, and lazy chains, Lovász and Kannan (1999) proved that

$$t_{\mathrm{mix}} \le 2000 \int_{\pi_{\min}}^{3/4} \frac{du}{u\Phi^2(u)}. \tag{17.44}$$

Morris and Peres (2005) sharpened this, using evolving sets, to obtain the following:

THEOREM (Morris and Peres (2005)). *For lazy irreducible Markov chains, if*

$$t \ge 1 + \int_{4(\pi(x)\wedge\pi(y))}^{4/\varepsilon} \frac{4du}{u\Phi^2(u)},$$

then

$$\left|\frac{P^t(x,y) - \pi(y)}{\pi(y)}\right| \le \varepsilon.$$

Note that this theorem does *not* require reversibility.

CHAPTER 18

The Cutoff Phenomenon

18.1. Definition

For the top-to-random shuffle on n cards, we obtained in Section 6.5.3 the bound

$$d_n(n \log n + \alpha n) \leq e^{-\alpha}, \tag{18.1}$$

while in Section 7.4.2 we showed that

$$\liminf_{n \to \infty} d_n(n \log n - \alpha n) \geq 1 - 2e^{2-\alpha}. \tag{18.2}$$

In particular, the upper bound in (18.1) tends to 0 as $\alpha \to \infty$, and the lower bound in (18.2) tends to 1 as $\alpha \to \infty$. It follows that $t_{\text{mix}}(\varepsilon) = n \log n \, [1 + h(n, \varepsilon)]$, where $\lim_{n \to \infty} h(n, \varepsilon) = 0$ for all ε. This is a much more precise statement than the fact that the mixing time is of the order $n \log n$.

The previous example motivates the following definition. Suppose, for a sequence of Markov chains indexed by $n = 1, 2, \ldots$, the mixing time for the n-th chain is denoted by $t_{\text{mix}}^{(n)}(\varepsilon)$. This sequence of chains has a ***cutoff*** if, for all $\varepsilon > 0$,

$$\lim_{n \to \infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1 - \varepsilon)} = 1. \tag{18.3}$$

The bounds (18.1) and (18.2) for the top-to-random chain show that the total variation distance d_n for the n-card chain "falls off a cliff" at $t_{\text{mix}}^{(n)}$. More precisely, when time is rescaled by $n \log n$, as $n \to \infty$ the function d_n approaches a step function:

$$\lim_{n \to \infty} d_n(cn \log n) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1. \end{cases} \tag{18.4}$$

In fact, this property characterizes when a sequence of chains has a cutoff.

LEMMA 18.1. *Let $t_{\text{mix}}^{(n)}$ and d_n be the mixing time and distance to stationarity, respectively, for the n-th chain in a sequence of Markov chains. The sequence has a cutoff if and only if*

$$\lim_{n \to \infty} d_n(ct_{\text{mix}}^{(n)}) = \begin{cases} 1 & \textit{if } c < 1, \\ 0 & \textit{if } c > 1. \end{cases}$$

The proof is left to the reader as Exercise 18.1.

Returning again to the example of the top-to-random shuffle on n cards, the bounds (18.1) and (18.2) show that in an interval of length αn centered at $n \log n$, the total variation distance decreased from near 1 to near 0. The next definition formalizes this property.

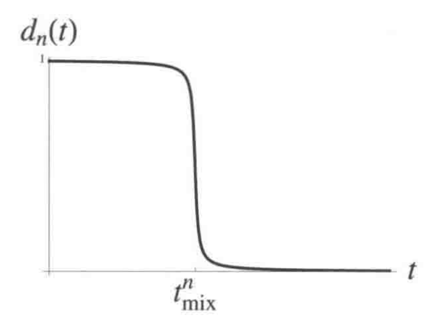


FIGURE 18.1. For a chain with a cutoff, the graph of $d_n(t)$ against t, when viewed on the time-scale of $t_{\text{mix}}^{(n)}$, approaches a step function as $n \to \infty$.

A sequence of Markov chains has a cutoff with a ***window*** of size $\{w_n\}$ if $w_n = o\left(t_{\text{mix}}^{(n)}\right)$ and

$$\lim_{\alpha \to -\infty} \liminf_{n \to \infty} d_n\left(t_{\text{mix}}^{(n)} + \alpha w_n\right) = 1,$$
$$\lim_{\alpha \to \infty} \limsup_{n \to \infty} d_n\left(t_{\text{mix}}^{(n)} + \alpha w_n\right) = 0.$$

We say a family of chains has a ***pre-cutoff*** if it satisfies the weaker condition

$$\sup_{0<\varepsilon<1/2} \limsup_{n \to \infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} < \infty.$$

Theorem 15.4 proved that the Glauber dynamics for the Ising model on the n-cycle has a pre-cutoff; it is an open problem to show that in fact this family of chains has a cutoff.

18.2. Examples of Cutoff

18.2.1. Biased random walk on a line segment. Let $p \in (1/2, 1)$ and $q = 1 - p$, so $\beta := (p - q)/2 = p - 1/2 > 0$. Consider the lazy nearest-neighbor random walk with bias β on the interval $\Omega = \{0, 1, \ldots, n\}$, which is the Markov chain with transition probabilities

$$P(k, k+1) = \begin{cases} \frac{p}{2} & \text{if } k \notin \{0, n\}, \\ \frac{1}{2} & \text{if } k = 0, \\ 0 & \text{if } k = n, \end{cases}$$
$$P(k, k) = \frac{1}{2},$$
$$P(k, k-1) = \begin{cases} \frac{q}{2} & \text{if } k \notin \{0, n\}, \\ 0 & \text{if } k = 0, \\ \frac{1}{2} & \text{if } k = n. \end{cases}$$

That is, when at an interior vertex, the walk remains in its current position with probability $1/2$, moves to the right with probability $p/2$, and moves to the left with probability $q/2$. When at an end-vertex, the walk remains in place with probability $1/2$ and moves to the adjacent interior vertex with probability $1/2$.

THEOREM 18.2. *The lazy random walk with bias $\beta = p - 1/2$ on $\{0, 1, 2, \ldots, n\}$ has a cutoff at $\beta^{-1}n$ with a window of order $\sqrt{n}$.*

PROOF. We write $t_n(\alpha) := \beta^{-1}n + \alpha\sqrt{n}$.

Upper bound, Step 1. We first prove that if $\tau_n := \min\{t \geq 0 : X_t = n\}$, then

$$\limsup_{n\to\infty} \mathbf{P}_0\{\tau_n > t_n(\alpha)\} \leq \Phi(-c(\beta)\alpha), \tag{18.5}$$

where $c(\beta)$ depends on β only and Φ is the standard normal distribution function.

Let (S_t) be a lazy β-biased nearest-neighbor random walk on all of $\mathbb{Z}$, so $\mathbf{E}_k S_t = k + \beta t$. We couple (X_t) to (S_t) until time $\tau_n := \min\{t \geq 0 : X_t = n\}$, as follows: let $X_0 = S_0$, and set

$$X_{t+1} = \begin{cases} 1 & \text{if } X_t = 0 \text{ and } S_{t+1} - S_t = -1, \\ X_t + (S_{t+1} - S_t) & \text{otherwise.} \end{cases} \tag{18.6}$$

This coupling satisfies $X_t \geq S_t$ for all $t \leq \tau_n$.

We have $\mathbf{E}_0 S_{t_n(\alpha)} = t_n(\alpha)\beta = n + \alpha\beta\sqrt{n}$, and

$$\mathbf{P}_0\{S_{t_n(\alpha)} < n\} = \mathbf{P}_0\left\{\frac{S_{t_n(\alpha)} - \mathbf{E}S_{t_n(\alpha)}}{\sqrt{t_n(\alpha)v}} < \frac{-\alpha\beta\sqrt{n}}{\sqrt{t_n(\alpha)v}}\right\},$$

where $v = 1/2 - \beta^2$. By the Central Limit Theorem, the right-hand side above converges as $n \to \infty$ to $\Phi(-c(\beta)\alpha)$. Thus

$$\limsup_{n\to\infty} \mathbf{P}_0\{S_{t_n(\alpha)} < n\} = \Phi(-c(\beta)\alpha). \tag{18.7}$$

Since $X_t \geq S_t$ for $t \leq \tau_n$,

$$\mathbf{P}_0\{\tau_n > t_n(\alpha)\} \leq \mathbf{P}_0\left\{\max_{0\leq s\leq t_n(\alpha)} S_s < n\right\} \leq \mathbf{P}_0\left\{S_{t_n(\alpha)} \leq n\right\},$$

which with (18.7) implies (18.5).

Upper bound, Step 2. We now show that we can couple two biased random walks so that the meeting time of the two walks is bounded by τ_n.

We couple as follows: toss a coin to decide which particle to move. Move the chosen particle up one unit with probability p and down one unit with probability q, unless it is at an end-vertex, in which case move it with probability one to the neighboring interior vertex. The time τ_{couple} until the particles meet is bounded by the time it takes the left-most particle to hit n, whence

$$d_n(t_n(\alpha)) \leq \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t_n(\alpha)\} \leq \mathbf{P}_0\{\tau_n > t_n(\alpha)\}.$$

This bound and (18.5) show that

$$\lim_{\alpha\to\infty} \limsup_{n\to\infty} d(t_n(\alpha)) \leq \lim_{\alpha\to\infty} \Phi(-c(\beta)\alpha) = 0.$$

Lower bound, Step 1. Let $\theta := (q/p)$. We first prove that

$$\limsup_{n\to\infty} \mathbf{P}_0\{X_{t_n(\alpha)} > n - h\} \leq 1 - \Phi(-c(\beta)\alpha) + \theta^{h-1}. \tag{18.8}$$

Let $(\tilde{X}_t)$ be the lazy biased random walk on $\{0, 1, \ldots\}$, with reflection at 0. By coupling with (X_t) so that $X_t \leq \tilde{X}_t$, for $x \geq 0$ we have

$$\mathbf{P}_0\{X_t > x\} \leq \mathbf{P}_0\{\tilde{X}_t > x\}. \tag{18.9}$$

Recall that (S_t) is the biased lazy walk on all of $\mathbb{Z}$. Couple $(\tilde{X}_t)$ with (S_t) so that $S_t \leq \tilde{X}_t$. Observe that $\tilde{X}_t - S_t$ increases (by a unit amount) only when $\tilde{X}_t$ is at 0, which implies that, for any t,

$$\mathbf{P}_0\{\tilde{X}_t - S_t \geq h\} \leq \mathbf{P}_0\{\text{at least } h-1 \text{ returns of } (\tilde{X}_t) \text{ to } 0\}.$$

By (9.21), the chance that the biased random walk on $\mathbb{Z}$, when starting from 1, hits 0 before n equals $1 - (1-\theta)/(1-\theta^n)$. Letting $n \to \infty$, the chance that the biased random walk on $\mathbb{Z}$, when starting from 1, ever visits 0 equals θ. Therefore,

$$\mathbf{P}_0\{\text{at least } h-1 \text{ returns of } (\tilde{X}_t) \text{ to } 0\} = \theta^{h-1},$$

and consequently,

$$\mathbf{P}_0\{\tilde{X}_t - S_t \geq h\} \leq \theta^{h-1}. \tag{18.10}$$

By (18.9) and (18.10),

$$\begin{aligned} \mathbf{P}_0\{X_{t_n(\alpha)} > n-h\} &\leq \mathbf{P}_0\{S_{t_n(\alpha)} > n-2h\} + \mathbf{P}_0\{\tilde{X}_{t_n(\alpha)} - S_{t_n(\alpha)} \geq h\} \\ &\leq \mathbf{P}_0\{S_{t_n(\alpha)} > n-2h\} + \theta^{h-1}. \end{aligned} \tag{18.11}$$

By the Central Limit Theorem,

$$\lim_{n\to\infty} \mathbf{P}_0\{S_{t_n(\alpha)} > n-2h\} = 1 - \Phi(-c(\beta)\alpha),$$

which together with (18.11) establishes (18.8).

Lower bound, Step 2. The stationary distribution equals

$$\pi^{(n)}(k) = \left[\frac{(p/q)-1}{(p/q)^{n+1}-1}\right](p/q)^k.$$

If $A_h = \{n-h+1, \ldots, n\}$, then

$$\pi^{(n)}(A_h) = \frac{1-(q/p)^{h+2}}{1-(q/p)^{n+1}}.$$

Therefore,

$$\begin{aligned} \liminf_{n\to\infty} d_n(t_n(\alpha)) &\geq \liminf_{n\to\infty} \left[\pi^{(n)}(A_h) - \mathbf{P}_0\{X_{t_n(\alpha)} > n-h\}\right] \\ &\geq 1 - \theta^{h+2} - \left[1 - \Phi(-c(\beta)\alpha) + \theta^{h-1}\right], \end{aligned}$$

and so

$$\lim_{\alpha\to-\infty} \liminf_{n\to\infty} d_n(t_n(\alpha)) \geq 1 - \theta^{h+2} - \theta^{h-1}.$$

Letting $h \to \infty$ shows that

$$\lim_{\alpha\to-\infty} \liminf_{n\to\infty} d_n(t_n(\alpha)) = 1.$$

■

18.2.2. Random walk on the hypercube. We return to the lazy random walk on the n-dimensional hypercube. In Section 5.3.3, it was shown that

$$t_{\mathrm{mix}}(\varepsilon) \leq n\log n + c_u(\varepsilon)n,$$

while Proposition 7.13 proved that

$$t_{\mathrm{mix}}(1-\varepsilon) \geq \frac{1}{2}n\log n - c_\ell(\varepsilon)n. \tag{18.12}$$

In fact, there is a cutoff, and the lower bound gives the correct constant:

THEOREM 18.3. *The lazy random walk on the n-dimensional hypercube has a cutoff at $(1/2)n \log n$ with a window of size n.*

PROOF. Let $\boldsymbol{X}_t = (X_t^1, \ldots, X_t^n)$ be the position of the random walk at time t, and let $W_t = W(\boldsymbol{X}_t) = \sum_{i=1}^n X_t^i$ be the Hamming weight of $\boldsymbol{X}_t$. As follows from the discussion in Section 2.3, (W_t) is a lazy version of the Ehrenfest urn chain whose transition matrix is given in (2.8). We write π_W for the stationary distribution of (W_t), which is binomial with parameters n and $1/2$.

The study of $(\boldsymbol{X}_t)$ can be reduced to the study of (W_t) because of the following identity:

$$\|\mathbf{P_1}\{\boldsymbol{X}_t \in \cdot\} - \pi\|_{\mathrm{TV}} = \|\mathbf{P}_n\{W_t \in \cdot\} - \pi_W\|_{\mathrm{TV}}. \tag{18.13}$$

Proof of (18.13). Let $\Omega_w := \{\boldsymbol{x} : W(\boldsymbol{x}) = w\}$. Note that by symmetry, the functions $\boldsymbol{x} \mapsto \mathbf{P_1}\{\boldsymbol{X}_t = \boldsymbol{x}\}$ and π are constant over Ω_w. Therefore,

$$\begin{aligned}\sum_{\boldsymbol{x} : W(\boldsymbol{x})=w} |\mathbf{P}_1\{\boldsymbol{X}_t = \boldsymbol{x}\} - \pi(\boldsymbol{x})| &= \left| \sum_{\boldsymbol{x} : W(\boldsymbol{x})=w} \mathbf{P}_1\{\boldsymbol{X}_t = \boldsymbol{x}\} - \pi(\boldsymbol{x}) \right| \\ &= |\mathbf{P}_1\{W_t = w\} - \pi_W(w)|.\end{aligned}$$

(The absolute values can be moved outside the sum in the first equality because all of the terms in the sum are equal.) Summing over $w \in \{0, 1, \ldots, n\}$ and dividing by 2 yields (18.13).

Since $(\boldsymbol{X}_t)$ is a transitive chain,

$$d(t) = \|\mathbf{P_1}\{\boldsymbol{X}_t \in \cdot\} - \pi\|_{\mathrm{TV}},$$

and it is enough to bound the right-hand side of (18.13).

We construct now a coupling (W_t, Z_t) of the lazy Ehrenfest chain started from w with the lazy Ehrenfest chain started from z. Provided that the two chains have not yet collided, at each move, a fair coin is tossed to determine which of the two chains moves; the chosen chain makes a transition according to the matrix (2.8), while the other chain remains in its current position. The chains move together once they have met for the first time.

Suppose, without loss of generality, that $z \geq w$. Since the chains never cross each other, $Z_t \geq W_t$ for all t. Consequently, if $D_t = |Z_t - W_t|$, then $D_t = Z_t - W_t \geq 0$. Let $\tau := \min\{t \geq 0 : Z_t = W_t\}$. Supposing that $(Z_t, W_t) = (z_t, w_t)$ and $\tau > t$,

$$D_{t+1} - D_t = \begin{cases} 1 & \text{with probability } (1/2)(1 - z_t/n) + (1/2)w_t/n, \\ -1 & \text{with probability } (1/2)z_t/n + (1/2)(1 - w_t/n). \end{cases} \tag{18.14}$$

From (18.14) we see that on the event $\{\tau > t\}$,

$$\mathbf{E}_{z,w}[D_{t+1} - D_t \mid Z_t = z_t, W_t = w_t] = -\frac{(z_t - w_t)}{n}. \tag{18.15}$$

Let $\boldsymbol{Z}_t = (Z_1, \ldots, Z_t)$ and $\boldsymbol{W}_t = (W_1, \ldots, W_t)$. By the Markov property and because $\mathbf{1}\{\tau > t\}$ is a function of $(\boldsymbol{Z}_t, \boldsymbol{W}_t)$,

$$\begin{aligned}&\mathbf{1}\{\tau > t\}\mathbf{E}_{z,w}[D_{t+1} - D_t \mid Z_t, W_t] \\ &\quad = \mathbf{1}\{\tau > t\}\mathbf{E}_{z,w}[D_{t+1} - D_t \mid \boldsymbol{Z}_t, \boldsymbol{W}_t] \\ &\quad = \mathbf{E}_{z,w}[\mathbf{1}\{\tau > t\}(D_{t+1} - D_t) \mid \boldsymbol{Z}_t, \boldsymbol{W}_t].\end{aligned} \tag{18.16}$$

Combining (18.15) and (18.16) shows that

$$\mathbf{E}_{z,w}[\mathbf{1}\{\tau > t\}D_{t+1} \mid \boldsymbol{Z}_t, \boldsymbol{W}_t] \leq \left(1 - \frac{1}{n}\right) D_t \mathbf{1}\{\tau > t\}.$$

Taking expectation, we have

$$\mathbf{E}_{z,w}[D_{t+1}\mathbf{1}\{\tau > t\}] = \left(1 - \frac{1}{n}\right) \mathbf{E}_{z,w}[D_t \mathbf{1}\{\tau > t\}].$$

Since $\mathbf{1}\{\tau > t+1\} \leq \mathbf{1}\{\tau > t\}$, we have

$$\mathbf{E}_{z,w}[D_{t+1}\mathbf{1}\{\tau > t+1\}] \leq \left(1 - \frac{1}{n}\right) \mathbf{E}_{z,w}[D_t \mathbf{1}\{\tau > t\}].$$

By induction,

$$\mathbf{E}_{z,w}[D_t \mathbf{1}\{\tau > t\}] \leq \left(1 - \frac{1}{n}\right)^t (z - w) \leq ne^{-t/n}. \tag{18.17}$$

Also, from (18.14), provided $\tau > t$, the process (D_t) is at least as likely to move downwards as it is to move upwards. Thus, until time τ, the process (D_t) can be coupled with a simple random walk (S_t) so that $S_0 = D_0$ and $D_t \leq S_t$.

If $\tilde{\tau} := \min\{t \geq 0 : S_t = 0\}$, then $\tau \leq \tilde{\tau}$. By Theorem 2.26, there is a constant c_1 such that for $k \geq 0$,

$$\mathbf{P}_k\{\tau > u\} \leq \mathbf{P}_k\{\tilde{\tau} > u\} \leq \frac{c_1 k}{\sqrt{u}}. \tag{18.18}$$

By (18.18),

$$\mathbf{P}_{z,w}\{\tau > s + u \mid D_0, D_1, \ldots D_s\} = \mathbf{1}\{\tau > s\}\mathbf{P}_{D_s}\{\tau > u\} \leq \frac{c_1 D_s \mathbf{1}\{\tau > s\}}{\sqrt{u}}.$$

Taking expectation above and applying (18.17) shows that

$$\mathbf{P}_{z,w}\{\tau > s + u\} \leq \frac{c_1 n e^{-s/n}}{\sqrt{u}}. \tag{18.19}$$

Letting $u = \alpha n$ and $s = (1/2)n \log n$ above, by Corollary 5.3 we have

$$d((1/2)n \log n + \alpha n) \leq \frac{c_1}{\sqrt{\alpha}}.$$

We conclude that

$$\lim_{\alpha \to \infty} \limsup_{n \to \infty} d((1/2)n \log n + \alpha n) = 0.$$

The lower bound (7.26) completes the proof. ■

18.3. A Necessary Condition for Cutoff

When does a family of chains have a cutoff? The following proposition gives a necessary condition.

PROPOSITION 18.4. *For a sequence of irreducible aperiodic Markov chains with relaxation times $\{t_{\mathrm{rel}}^{(n)}\}$ and mixing times $\{t_{\mathrm{mix}}^{(n)}\}$, if $t_{\mathrm{mix}}^{(n)}/t_{\mathrm{rel}}^{(n)}$ is bounded above, then there is no pre-cutoff.*

PROOF. The proof follows from Theorem 12.4: dividing both sides of (12.12) by $t_{\mathrm{mix}}^{(n)}$, we have

$$\frac{t_{\mathrm{mix}}^{(n)}(\varepsilon)}{t_{\mathrm{mix}}^{(n)}} \geq \frac{t_{\mathrm{rel}}^{(n)} - 1}{t_{\mathrm{mix}}^{(n)}} \log\left(\frac{1}{2\varepsilon}\right) \geq c_1 \log\left(\frac{1}{2\varepsilon}\right).$$

As $\varepsilon \to 0$, the right-hand side increases to infinity. ■

Recall that we write $a_n \asymp b_n$ to mean that there exist positive and finite constants c_1 and c_2, not depending on n, such that $c_1 \leq a_n/b_n \leq c_2$ for all n.

EXAMPLE 18.5. Consider the lazy random walk on the cycle $\mathbb{Z}_n$. In Section 5.3.1 we showed that $t_{\mathrm{mix}}^{(n)} \leq n^2$. In fact, this is the correct order, as shown in Section 7.4.1. In Section 12.3.1, we computed the eigenvalues of the transition matrix, finding that $t_{\mathrm{rel}}^{(n)} \asymp n^2$ also. By Proposition 18.4, there is no pre-cutoff.

EXAMPLE 18.6. Let T_n be the rooted binary tree with n vertices. In Example 7.7, we showed that the lazy simple random walk has $t_{\mathrm{mix}} \asymp n$. Together with Theorem 12.4, this implies that there exists a constant c_1 such that $t_{\mathrm{rel}} \leq c_1 n$. In Example 7.7, we actually showed that $\Phi_\star \leq 1/(n-2)$. Thus, by Theorem 13.14, we have $\gamma \leq 2/(n-2)$, whence $t_{\mathrm{rel}} \geq c_2 n$ for some constant c_2. An application of Proposition 18.4 shows that there is no pre-cutoff for this family of chains.

The question remains if there are conditions which ensure that the converse of Proposition 18.4 holds. Below we give a variant of an example due to Igor Pak (personal communication) which shows the converse is not true in general.

EXAMPLE 18.7. Let $\{P_n\}$ be a family of transition matrices with $t_{\mathrm{rel}}^{(n)} = o(t_{\mathrm{mix}}^{(n)})$ and with a cutoff (e.g., the lazy random walk on the hypercube.) Let $L_n := \sqrt{t_{\mathrm{rel}}^{(n)} t_{\mathrm{mix}}^{(n)}}$, and define the matrix

$$\tilde{P}_n = (1 - 1/L_n)P_n + (1/L)\Pi_n,$$

where $\Pi_n(x, y) := \pi_n(y)$ for all x.

We first prove that

$$\left\|\tilde{P}_n^t(x, \cdot) - \pi\right\|_{\mathrm{TV}} = \left(1 - \frac{1}{L_n}\right)^t \left\|P_n^t(x, \cdot) - \pi\right\|_{\mathrm{TV}}. \tag{18.20}$$

Proof of (18.20). One step of the chain can be generated by first tossing a coin with probability $1/L_n$ of heads; if heads, a sample from π_n is produced, and if tails, a transition from P_n is used. If τ is the first time that the coin lands heads, then τ has a geometric distribution with success probability $1/L_n$. Accordingly,

$$\begin{aligned}
\mathbf{P}_x\{X_t^{(n)} = y\} - \pi(y) &= \mathbf{P}_x\{X_t^{(n)} = y, \tau \leq t\} + \mathbf{P}_x\{X_t^{(n)} = y, \tau > t\} - \pi(y) \\
&= -\pi(y)[1 - \mathbf{P}_x\{\tau \leq t\}] + P_n^t(x, y)\mathbf{P}_x\{\tau > t\} \\
&= \left[P_n^t(x, y) - \pi_n(y)\right] \mathbf{P}_x\{\tau > t\}.
\end{aligned}$$

Taking absolute value and summing over y gives (18.20). We conclude that

$$\tilde{d}_n(t) = (1 - L_n^{-1})^t d_n(t).$$

Therefore,

$$\tilde{d}_n(\beta L_n) \leq e^{-\beta} d_n(\beta L_n) \leq e^{-\beta},$$

and $\tilde{t}_{\text{mix}}^{(n)} \leq c_1 L_n$ for some constant c_1. On the other hand

$$\tilde{d}_n(\beta L_n) = e^{-\beta[1+o(1)]} d_n(\beta L_n). \tag{18.21}$$

Since $L_n = o(t_{\text{mix}}^{(n)})$ and the P_n-chains have a cutoff, we have that $d_n(\beta L_n) \to 1$ for all β, whence from the above,

$$\lim_{n\to\infty} \tilde{d}_n(\beta L_n) = e^{-\beta}.$$

This shows both that $\tilde{t}_{\text{mix}}^{(n)} \asymp L_n$ and that there is no pre-cutoff for the $\tilde{P}$-chains.

Let $\{\lambda_j^{(n)}\}_{j=1}^n$ be the eigenvalues of P_n. As can be directly verified, $\lambda_1^{(n)}$ is an eigenvalue of $\tilde{P}_n$, and $\tilde{\lambda}_j^{(n)} := (1-1/L_n)\lambda_j^{(n)}$ is an eigenvalue of $\tilde{P}_n$ for $j > 1$. Thus,

$$\tilde{\gamma}_n = 1 - \left(1 - \frac{1}{L_n}\right)(1-\gamma_n) = \gamma_n\left[1 + o(1)\right].$$

We conclude that $\tilde{t}_{\text{rel}}^{(n)} \asymp t_{\text{rel}}^{(n)}$. However, $\tilde{t}_{\text{rel}}^{(n)} = o(\tilde{t}_{\text{mix}})$, since $\tilde{t}_{\text{mix}} \asymp L_n$.

18.4. Separation Cutoff

The mixing time can be defined for other distances. The separation distance, defined in Section 6.4, is $s(t) = \max_{x\in\Omega} s_x(t)$, where

$$s_x(t) := \max_{y\in\Omega}\left[1 - \frac{P^t(x,y)}{\pi(y)}\right].$$

We define

$$t_{\text{sep}}(\varepsilon) := \inf\{t \geq 0 : s(t) \leq \varepsilon\}.$$

A family of Markov chains with separation mixing times $\{t_{\text{sep}}^{(n)}\}$ has a ***separation cutoff*** if

$$\lim_{n\to\infty} \frac{t_{\text{sep}}^{(n)}(\varepsilon)}{t_{\text{sep}}^{(n)}(1-\varepsilon)} = 1 \quad \text{for all } \varepsilon > 0.$$

THEOREM 18.8. *The lazy random walk on the n-dimensional hypercube has a separation cutoff at $n \log n$ with a window of order n.*

PROOF. We already have proven a sufficient upper bound in Section 6.5.2:

$$s(n\log n + \alpha n) \leq e^{-\alpha}. \tag{18.22}$$

We are left with the task of proving a lower bound. Recall that τ_{refresh} is the strong stationary time equal to the first time all the coordinates have been selected for updating. Since, when starting from $\mathbf{1}$, the state $\mathbf{0}$ is a halting state for τ_{refresh}, it follows that

$$s_{\mathbf{1}}(t) = \mathbf{P}_{\mathbf{1}}\{\tau_{\text{refresh}} > t\}.$$

(See Remark 6.12.)

Let R_t be the number of coordinates not updated by time t. Let $t_n := n\log n - \alpha n$. By Lemma 7.12, we have

$$\mathbf{E}R_{t_n} = n(1-n^{-1})^{t_n} \to e^{\alpha} \quad \text{and} \quad \text{Var}(R_{t_n}) \leq e^{\alpha}.$$

Therefore,

$$\mathbf{P}_{\mathbf{1}}\{\tau_{\text{refresh}} \leq t_n\} = \mathbf{P}_{\mathbf{1}}\{R_{t_n} = 0\} \leq c_1 e^{-\alpha},$$

for some constant c_1. Thus,

$$s_{\mathbf{1}}(n\log n - \alpha n) \geq 1 - c_1 e^{-\alpha}. \tag{18.23}$$

The bounds (18.22) and (18.23) together imply a separation cutoff at $n \log n$ with a window of size n. ■

Exercise

EXERCISE 18.1. Let $t^n_{\rm mix}$ and d_n denote the mixing time and distance to stationarity, respectively, for the n-th chain in a sequence of Markov chains. Show that the sequence has a cutoff if and only if

$$\lim_{n\to\infty} d_n(ct^n_{\rm mix}) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 0. \end{cases} \tag{18.24}$$

Notes

The biased random walk on the interval is studied in Diaconis and Fill (1990); see also the discussion in Diaconis and Saloff-Coste (2006), which contains many examples. More on cutoff is discussed in Chen and Saloff-Coste (2008).

A chain with pre-cutoff, but no cutoff. David Aldous (2004) created the chain whose transition probabilities are shown in Figure 18.2. The shape of the graph of $d(t)$ as a function of t is shown on the bottom of the figure. Since the stationary distribution grows geometrically from left-to-right, the chain mixes once it reaches near the right-most point. It takes about $15n$ steps for a particle started at the left-most endpoint to reach the fork. With probability about $3/4$, it first reaches the right endpoint via the bottom path. (This can be calculated using effective resistances; cf. Section 9.4.) When the walker takes the bottom path, it takes about $(5/3)n$ additional steps to reach the right. In fact, the time will be within order $\sqrt{n}$ of $(5/3)n$ with high probability. In the event that the walker takes the top path, it takes about $6n$ steps (again $\pm O(\sqrt{n})$) to reach the right endpoint. Thus the total variation distance will drop by $3/4$ at time $[15 + (5/3)]n$, and it will drop by the remaining $1/4$ at around time $(15+6)n$. Both of these drops will occur within windows of order $\sqrt{n}$. Thus, the ratio $t_{\rm mix}(\varepsilon)/t_{\rm mix}(1-\varepsilon)$ will stay bounded as $n \to \infty$, but it does not tend to 1.

Recently, Lubetzky and Sly (2008) have announced a proof of cutoff for random regular graphs:

THEOREM (Lubetzky and Sly (2008)). *Let G be a random d-regular graph for $d \geq 3$ fixed. Then with high probability, the simple random walk on G exhibits cutoff at time $\frac{d}{d-2} \log_{d-1} n$ with a window of order $\sqrt{\log n}$.*

Ding, Lubetzky, and Peres (2008b) analyzed the cutoff phenomena for birth-and-death chains. They proved

THEOREM (Ding et al. (2008b)). *For any $0 < \varepsilon < \frac{1}{2}$ there exists an explicit $c_\varepsilon > 0$ such that every lazy irreducible birth-and-death chain (X_t) satisfies*

$$t_{\rm mix}(\varepsilon) - t_{\rm mix}(1-\varepsilon) \leq c_\varepsilon \sqrt{t_{\rm rel} \cdot t_{\rm mix}(\tfrac{1}{4})}. \tag{18.25}$$

COROLLARY (Ding et al. (2008b)). *Let $(X^{(n)}_t)$ be a sequence of lazy irreducible birth-and-death chains. Then it exhibits cutoff in total-variation distance if and only if $t^{(n)}_{\rm mix} \cdot \gamma(n)$ tends to infinity with n. Furthermore, the cutoff window size is at most the geometric mean between the mixing time and relaxation time.*

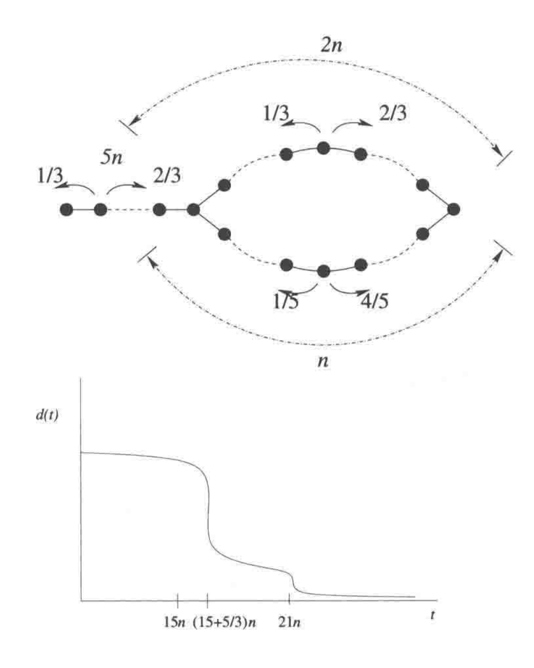


FIGURE 18.2. Random walk on the network shown on the top has a pre-cutoff, but no cutoff. The shape of the graph of $d(t)$ is shown on the bottom.

Earlier, Diaconis and Saloff-Coste (2006) obtained a similar result for separation cutoff.

CHAPTER 19

Lamplighter Walks

19.1. Introduction

Imagine placing a lamp at each vertex of a finite graph $G = (V, E)$. Now allow a (possibly intoxicated?) lamplighter to perform a random walk on G, switching lights randomly on and off as he visits them.

This process can be modeled as a random walk on the ***wreath product*** $G^* = \{0,1\}^V \times V$, whose vertices are ordered pairs (f, v) with $v \in V$ and $f \in \{0,1\}^V$. There is an edge between (f, v) and (h, w) in the graph G^* if v, w are adjacent or identical in G and $f(u) = h(u)$ for $u \notin \{v, w\}$. We call f the ***configuration of the lamps*** and v the ***position of the lamplighter***. In the configuration function f, zeroes correspond to lamps that are off, and ones correspond to lamps that are on.

We now construct a Markov chain on G^*. Let Υ denote the transition matrix for the lamplighter walk, and let P be the transition matrix of the lazy simple random walk on G.

- For $v \neq w$, $\Upsilon[(f, v), (h, w)] = P(v, w)/4$ if f and h agree outside of $\{v, w\}$.
- When $v = w$, $\Upsilon[(f, v), (h, v)] = P(v, v)/2$ if f and h agree off of $\{v\}$.

That is, at each time step, the current lamp is randomized, the lamplighter moves, and then the new lamp is also randomized. (The lamp at w is randomized in order to make the chain reversible. We have used the lazy walk on G as the basis for the construction to avoid periodicity problems later.) We will assume throughout this chapter that G is connected, which implies that both P and Υ are irreducible. We write π for the stationary distribution of P, and π^* for the stationary distribution of Υ.

Since the configuration of lamps on visited states is uniformly distributed, allowing the lamplighter to walk for the cover time of the underlying walk suffices to

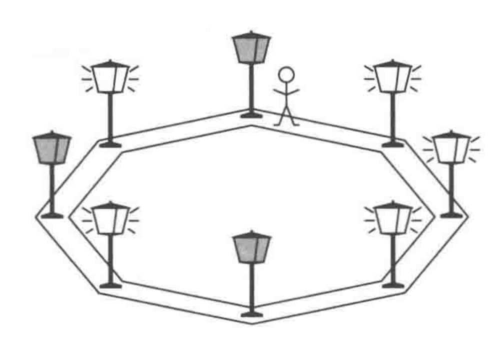


FIGURE 19.1. A lamplighter on an 8-cycle.

randomize the lamp configuration—although perhaps not the position of the lamplighter himself! Does the lamplighter need to walk further to mix? In this chapter we study several connections between the underlying chain G and the lamplighter chain G^*.

We have by now defined several time parameters associated with a finite Markov chain. Some measure mixing directly; others, such as the cover time and the hitting time, attempt to measure the geometry of the chain. Below, in (19.1), we summarize some inequalities we have proved relating these parameters. Define $t_1 \lesssim t_2$ if there exists a constant $c > 0$ such that $t_1 \lesssim ct_2$. We have shown

$$t_{\text{rel}} \lesssim t_{\text{mix}} \lesssim t_{\text{hit}} \lesssim t_{\text{cov}}, \tag{19.1}$$

where the first inequality hold for reversible chains (Theorem 12.4), the second inequality holds for reversible and lazy chains (Remark 10.16), and the last holds generally (see Equation (11.2)).

In the next section, we prove that the relaxation time t_{rel} of the lamplighter walk is comparable to the maximal hitting time t_{hit} of the underlying walk (Theorem 19.1). In Section 19.3, we show (Theorem 19.2) that the cover time t_{cov} of the walk on G is comparable to the mixing time for the lamplighter walk on G^*. The proofs of these results use many of the techniques we have studied in previous chapters.

19.2. Relaxation Time Bounds

THEOREM 19.1. *For each n, let G_n be a graph with vertex set V_n, and suppose that $|V_n| \to \infty$. Then there exist constants $c_1 < c_2$ such that for sufficiently large n,*

$$c_1 t_{\text{hit}}(G_n) \leq t_{\text{rel}}(G_n^*) \leq c_2 t_{\text{hit}}(G_n). \tag{19.2}$$

PROOF OF THEOREM 19.1. To prove the lower bound, we use the variational formula of Lemma 13.12 to show that the spectral gap for the transition matrix Υ^t is bounded away from 1 when $t = t_{\text{hit}}(G_n)/4$. For the upper bound, we use the coupling contraction method of Chen (1998), which we have already discussed (Theorem 13.1). The geometry of lamplighter graphs allows us to refine this coupling argument and restrict our attention to pairs of states such that the position of the lamplighter is the same in both states.

Lower bound. Fix a vertex $w \in G$ that maximizes $\mathbf{E}_\pi(\tau_w)$, and define $\varphi : V^* \to \{0,1\}$ by $\varphi(f,v) = f(w)$. Then $\text{Var}_{\pi^*}(\varphi) = 1/4$. Let (Y_t) be the Markov chain on G^* with initial distribution π^*, so that Y_t has distribution π^* for all $t \geq 0$. We write $Y_t = (F_t, X_t)$, where X_t is the position of the walk at time t, and F_t is the configuration of lamps at time t. Applying Lemma 13.11 to Υ^t and then conditioning on the walk's path up to time t shows that

$$\begin{aligned} \mathcal{E}_t(\varphi) &= \frac{1}{2}\mathbf{E}_{\pi^*}\left[\varphi(Y_t) - \varphi(Y_0)\right]^2 \\ &= \frac{1}{2}\mathbf{E}_{\pi^*}\left(\mathbf{E}_{\pi^*}[(\varphi(Y_t) - \varphi(Y_0))^2 \mid X_0, \ldots, X_t]\right). \end{aligned} \tag{19.3}$$

Observe that

$$\begin{aligned} \mathbf{E}_{\pi^*}[(\varphi(Y_t) - \varphi(Y_0))^2 \mid X_0, \ldots X_t] &= \mathbf{E}_{\pi^*}[F_t(w) - F_0(w) \mid X_0, \ldots, X_t] \\ &= \frac{1}{2}\mathbf{1}_{\{\tau_w \leq t\}}, \end{aligned}$$

as $F_t(w) - F_0(w) = 1$ if and only if the walk visits w by time t, and, at the walk's last visit to w before or at time t, the lamp at w is refreshed to a state different from its initial state. Combining the above equality with (19.3) shows that

$$\mathcal{E}_t(\varphi) = \frac{1}{4}\mathbf{P}_\pi\{\tau_w \le t\}. \tag{19.4}$$

For any t,

$$\mathbf{E}_v\tau_w \le t + t_{\text{hit}}\mathbf{P}_v\{\tau_w > t\}. \tag{19.5}$$

This follows because if a walk on G started at v has not hit w by time t, the expected additional time to arrive at w is bounded by t_{hit}, regardless of the value of the state at time t. Averaging (19.5) over π shows that

$$\mathbf{E}_\pi\tau_w \le t + t_{\text{hit}}\mathbf{P}_\pi\{\tau_w > t\}. \tag{19.6}$$

By Lemma 10.2 and our choice of w, we have $t_{\text{hit}} \le 2\mathbf{E}_\pi\tau_w$, whence (19.6) implies that

$$t_{\text{hit}} \le 2t + 2t_{\text{hit}}\mathbf{P}_\pi\{\tau_w > t\}.$$

Substituting $t = t_{\text{hit}}/4$ and rearranging yields

$$\mathbf{P}_\pi\{\tau_w \le t_{\text{hit}}/4\} \le \frac{3}{4}.$$

By Remark 13.13 and (19.4), we thus have

$$1 - \lambda_2^{t_{\text{hit}}/4} \le \frac{3}{4}.$$

Therefore

$$\log 4 \ge \frac{t_{\text{hit}}}{4}(1-\lambda_2),$$

which gives the claimed lower bound on $t_{\text{rel}}(G^*)$, with $c_1 = \frac{1}{\log 4}$. (Note that since the walk is lazy, $|\lambda_2| = \lambda_2$.)

Upper bound. We use a coupling argument related to that of Theorem 13.1. Suppose that φ is an eigenfunction for Υ with eigenvalue λ_2. To conclude that $t_{\text{rel}}(G^*) \le \frac{(2+o(1))t_{\text{hit}}}{\log 2}$, it suffices to show that $\lambda_2^{2t_{\text{hit}}+o(1)} \le 1/2$. Note that for lamp configurations f and g on G, the ℓ^1 norm $\|f-g\|_1$ is equal to the number of bits in which f and g differ. Let

$$M = \max_{f,g,x} \frac{|\varphi(f,x) - \varphi(g,x)|}{\|f-g\|_1}.$$

(Note that M is a restricted version of a Lipschitz constant: the maximum is taken only over states with the same lamplighter position.)

If $M=0$, then $\varphi(f,x)$ depends only on x and $\psi(x) = \varphi(f,x)$ is an eigenfunction for the transition matrix P. Theorem 12.4 and (10.24) together imply that

$$(\log 2)t_{\text{rel}} \le [2+o(1)]t_{\text{hit}}.$$

Hence

$$\left|\lambda_2^{(2+o(1))t_{\text{hit}}}\right| \le \left|\lambda_2^{(\log 2)(1/(1-\lambda_2))}\right| \le \frac{1}{2},$$

since $\sup_{x\in[0,1)} x^{(\log 2)/(1-x)} = 1/2$. We may thus assume that $M > 0$.

Couple two lamplighter walks, one started at (f,x) and one at (g,x), by using the same lamplighter steps and updating the configurations so that they agree at each site visited by the lamplighter. Let (f',x') and (g',x') denote the positions

of the coupled walks after $2t_{\text{hit}}$ steps, and let K denote the combined transition matrix of this coupling. Because φ is an eigenfunction for Υ,

$$\begin{aligned}
\lambda_2^{2t_{\text{hit}}} M &= \sup_{f,g,x} \frac{|\Upsilon^{2t_{\text{hit}}}\varphi(f,x) - \Upsilon^{2t_{\text{hit}}}\varphi(g,x)|}{\|f-g\|_1} \\
&\le \sup_{f,g,x} \sum_{f',g',x'} K^{2t_{\text{hit}}}[(f,g,x),(f',g',x')] \frac{|\varphi(f',x') - \varphi(g',x')|}{\|f'-g'\|_1} \frac{\|f'-g'\|_1}{\|f-g\|_1} \\
&\le M \sup_{f,g,x} \frac{\mathbf{E}\,\|f'-g'\|_1}{\|f-g\|_1}.
\end{aligned}$$

But at time $2t_{\text{hit}}$, each lamp that contributes to $\|f-g\|_1$ has probability of at least $1/2$ of having been visited, and so $\mathbf{E}\,\|f'-g'\|_1 \le \|f-g\|_1/2$. Dividing by M gives the required bound of $\lambda_2^{2t_{\text{hit}}} \le 1/2$. ■

19.3. Mixing Time Bounds

THEOREM 19.2. *Let (G_n) be a sequence of graphs with $|V_n| \to \infty$, and let $t_{\text{cov}}^{(n)}$ be the cover time for lazy simple random walk on G_n. There exist constants c_1 and c_2 such that for sufficiently large n, the mixing times of the lamplighter family (G_n^*) satisfy*

$$c_1 t_{\text{cov}}^{(n)} \le t_{\text{mix}}(G_n^*) \le c_2 t_{\text{cov}}^{(n)}. \tag{19.7}$$

We first prove three lemmas needed in the proof of the lower bound.

Aldous and Diaconis (1987) proved the following inequality. Recall the definitions (6.7) and (4.23) of s and $\bar{d}$, respectively.

LEMMA 19.3. *For a reversible chain, the separation and total variation distances satisfy*

$$s(2t) \le 1 - (1 - \bar{d}(t))^2. \tag{19.8}$$

Proof of (19.8). By reversibility, $P^t(z,y)/\pi(y) = P^t(y,z)/\pi(z)$, whence

$$\frac{P^{2t}(x,y)}{\pi(y)} = \sum_{z\in\Omega} \frac{P^t(x,z)P^t(z,y)}{\pi(y)} = \sum_{z\in\Omega} \pi(z) \frac{P^t(x,z)P^t(y,z)}{\pi(z)^2}.$$

Applying Cauchy-Schwarz to the right-hand side above, we have

$$\begin{aligned}
\frac{P^{2t}(x,y)}{\pi(y)} &\ge \left(\sum_{z\in\Omega} \sqrt{P^t(x,z)P^t(y,z)} \right)^2 \\
&\ge \left(\sum_{z\in\Omega} P^t(x,z) \wedge P^t(y,z) \right)^2.
\end{aligned}$$

From equation (4.13),

$$\frac{P^{2t}(x,y)}{\pi(y)} \ge \left(1 - \|P^t(x,\cdot) - P^t(y,\cdot)\|_{\text{TV}}\right)^2 \ge \left(1 - \bar{d}(t)\right)^2.$$

Subtracting both sides of the inequality from 1 and maximizing over x and y yields (19.8).

LEMMA 19.4. *For the lamplighter chain G^* on a finite graph G with vertex set V having $|V| = n$, the separation distance $s^*(t)$ satisfies*

$$s^*(t) \geq \mathbf{P}_w\{\tau_{\text{cov}} > t\} \tag{19.9}$$

for every $w \in V$ and $t > 0$.

PROOF. Let w_t be the vertex minimizing $\mathbf{P}_w\{X_t = w_t \mid \tau_{\text{cov}} \leq t\}/\pi(w_t)$. Since $\mathbf{P}_w\{X_t = \cdot \mid \tau_{\text{cov}} \leq t\}$ and π are both probability distributions on V, we have $\mathbf{P}_w\{X_t = w_t \mid \tau_{\text{cov}} \leq t\} \leq \pi(w_t)$. Since the only way to go from all lamps off to all lamps on is to visit every vertex, we have

$$\begin{aligned}\frac{\Upsilon^t((\mathbf{0}, w), (\mathbf{1}, w_t))}{\pi^*(\mathbf{1}, w_t)} &= \frac{\mathbf{P}_w\{\tau_{\text{cov}} \leq t\}2^{-n}\mathbf{P}_w\{X_t = w_t \mid \tau_{\text{cov}} \leq t\}}{2^{-n}\pi(w_t)} \\ &\leq \mathbf{P}_w\{\tau_{\text{cov}} \leq t\}.\end{aligned} \tag{19.10}$$

Subtracting from 1 yields $s^*(t) \geq \mathbf{P}_w\{\tau_{\text{cov}} > t\}$. ■

LEMMA 19.5. *Consider an irreducible finite Markov chain on state space Ω with transition matrix P, and let τ_{cov} be its cover time. Let t_m have the following property: for any $x \in \Omega$,*

$$\mathbf{P}_x\{\tau_{\text{cov}} \leq t_m\} \geq 1/2.$$

Then $t_{\text{cov}} \leq 2t_m$.

PROOF. Consider starting at a state $x \in \Omega$ and running in successive intervals of t_m steps. The probability of states being missed in the first interval is at most 1/2. If some states are missed in the first interval, then the probability that all are covered by the end of the second interval is at least 1/2, by the definition of t_m. Hence the probability of not covering by time $2t_m$ is at most 1/4. In general,

$$\mathbf{P}_x\{\tau_{\text{cov}} > kt_m\} \leq \frac{1}{2^k}.$$

We may conclude that τ_{cov} is dominated by t_m times a geometric random variable with success probability 1/2, and thus t_{cov} is at most $2t_m$. ■

PROOF OF THEOREM 19.2. Note: throughout the proof, asterisks indicate parameters for the lamplighter chain.

Upper bound. Let (F_t, X_t) denote the state of the lamplighter chain at time t. We will run the lamplighter chain long enough that, with high probability, every lamp has been visited and enough additional steps have been taken to randomize the position of the lamplighter.

Set $u_n = 8t_{\text{cov}}^{(n)} + t_{\text{mix}}(G_n, 1/8)$ and fix an initial state $(\mathbf{0}, v)$. Define the probability distribution $\mu_s^{u_n}$ on G_n^* by

$$\mu_s^{u_n} = \mathbf{P}_{(\mathbf{0},v)}\{(F_{u_n}, X_{u_n}) \in \cdot \mid \tau_{\text{cov}}^{(n)} = s\}.$$

Then

$$\Upsilon^{u_n}((0, v), \cdot) = \sum_s \mathbf{P}_{(\mathbf{0},v)}\{\tau_{\text{cov}}^{(n)} = s\}\mu_s^{u_n}.$$

By the triangle inequality,

$$\|\Upsilon^{u_n}((\mathbf{0}, v), \cdot) - \pi^*\|_{TV} \leq \sum_s \mathbf{P}_{(\mathbf{0},v)}\{\tau_{\text{cov}}^{(n)} = s\} \|\mu_s^{u_n} - \pi^*\|_{TV}. \tag{19.11}$$

Since $\mathbf{P}\{\tau_{\text{cov}}^{(n)} > 8t_{\text{cov}}^{(n)}\} < 1/8$ and the total variation distance between distributions is bounded by 1, we can bound

$$\|\Upsilon^{u_n}((\mathbf{0}, v), \cdot) - \pi^*\|_{TV} \leq 1/8 + \sum_{s \leq 8t_{\text{cov}}^{(n)}} \mathbf{P}_{(\mathbf{0},v)}\{\tau_{\text{cov}}^{(n)} = s\} \|\mu_s^{u_n} - \pi^*\|_{TV}. \quad (19.12)$$

Let ν_n denote the uniform distribution on $\{0,1\}^n$. For $s \leq u_n$, conditional on $\tau_{\text{cov}}^{(n)} = s$ and $X_s = x$, the distribution of F_{u_n} equals ν_n, the distribution of X_{u_n} is $P^{u_n - s}(x, \cdot)$, and F_{u_n} and X_{u_n} are independent. Thus,

$$\begin{aligned}
\mu_s^{u_n} &= \sum_{x \in V} \mathbf{P}_{(\mathbf{0},v)}\{(F_{u_n}, X_{u_n}) \in \cdot \mid \tau_{\text{cov}}^{(n)} = s,\, X_s = x\} \mathbf{P}_{(\mathbf{0},v)}\{X_s = x \mid \tau_{\text{cov}}^{(n)} = s\} \\
&= \sum_{x \in V} [\nu_n \times P^{u_n - s}(x, \cdot)] \mathbf{P}_{(\mathbf{0},v)}\{X_s = x \mid \tau_{\text{cov}}^{(n)} = s\}.
\end{aligned}$$

By the triangle inequality and Exercise 4.5, since $\pi^\star = \nu_n \times \pi$,

$$\begin{aligned}
\|\mu_s^{u_n} - \pi^\star\|_{TV} &\leq \sum_{x \in V} \left\|\nu_n \times P^{u_n - s}(x, \cdot) - \pi^\star\right\|_{TV} \mathbf{P}_{(\mathbf{0},v)}\{X_s = x \mid \tau_{\text{cov}}^{(n)} = s\} \\
&\leq \max_{x \in V} \left\|P^{u_n - s}(x, \cdot) - \pi\right\|_{TV}. \quad (19.13)
\end{aligned}$$

For $s \leq 8t_{\text{cov}}^{(n)}$, we have $u_n - s \geq t_{\text{mix}}(G_n, 1/8)$, by definition of u_n. Consequently, by (19.13), for $s \leq 8t_{\text{cov}}^{(n)}$,

$$\|\mu_s^{u_n} - \pi^\star\|_{TV} \leq \frac{1}{8}. \quad (19.14)$$

Using (19.14) in (19.12) shows that

$$\|\Upsilon^{u_n}((\mathbf{0}, v), \cdot) - \pi^*\|_{TV} \leq 1/8 + (1)(1/8) = 1/4. \quad (19.15)$$

To complete the upper bound, we need only recall from (19.1) that $t_{\text{mix}}(G_n, 1/8)$ is bounded by a constant times $t_{\text{cov}}^{(n)}$.

Lower bound. Lemmas 4.11 and 4.12 imply that

$$\bar{d}^*(2t_{\text{mix}}^*) \leq 1/4,$$

and Lemma 19.3 yields

$$s^*(4t_{\text{mix}}^*) \leq 1 - (1 - \bar{d}^*(2t_{\text{mix}}^*))^2 \leq 1 - (3/4)^2 < 1/2.$$

By Lemma 19.4 applied to G_n with $t = 4t_{\text{mix}}^*$, we have

$$\mathbf{P}_w\{\tau_{\text{cov}}^{(n)} > t_{\text{mix}}^*\} < 1/2.$$

Lemma 19.5 now immediately implies $t_{\text{cov}}^{(n)} \leq 8t_{\text{mix}}^*$. ■

19.4. Examples

19.4.1. The complete graph. When G_n is the complete graph on n vertices, with self-loops, then the chain we study on G_n^* is a random walk on the hypercube—although not quite the standard one, since two bits can change in a single step. This example was analyzed by Häggström and Jonasson (1997). The maximal hitting time is n and the expected cover time is an example of the coupon collector problem. Hence the relaxation time and the mixing time for G_n^* are $\Theta(n)$ and $\Theta(n \log n)$, respectively, just as for the standard walk on the hypercube.

19.4.2. Hypercube. Let $G_n = \mathbb{Z}_2^n$, the n-dimensional hypercube. We showed in Exercise 10.5 that the maximal hitting time is on the order of 2^n and in Exercise 11.3 that the cover time is on the order of $n2^n$. In Example 12.15, we saw that for lazy random walk on G_n, we have $t_{\text{rel}}(G_n) = n$. Finally, in Section 12.5, we showed that $t_{\text{mix}}(\varepsilon, G_n) \sim (n \log n)/2$. By Theorem 19.1, $t_{\text{rel}}(G_n^*)$ is on the order of 2^n, and Theorem 19.2 shows that the convergence time in total variation on G_n^* is on the order of $n2^n$.

19.4.3. Tori. For the one-dimensional case, we note that Häggström and Jonasson (1997) examined lamplighter walks on cycles. Here both the maximal hitting time and the expected cover time of the base graph are $\Theta(n^2)$—see Section 2.1 and Example 11.1. Hence the lamplighter chain on the cycle has both its relaxation time and its mixing time of order $\Theta(n^2)$.

For higher-dimensional tori, we have proved enough about hitting and cover times to see that the relaxation time and the mixing time grow at different rates in every dimension $d \geq 2$.

THEOREM 19.6. *For the random walk (X_t) on $(\mathbb{Z}_n^2)^*$ in which the lamplighter performs simple random walk with holding probability $1/2$ on $\mathbb{Z}_n^2$, there exist constants c_2 and C_2 such that the relaxation time satisfies*

$$c_2 n^2 \log n \leq t_{\text{rel}}((\mathbb{Z}_n^2)^*) \leq C_2 n^2 \log n. \tag{19.16}$$

There also exist constants c_2' and C_2' such that the total variation mixing time satisfies

$$c_2' n^2 (\log n)^2 \leq t_{\text{mix}}((\mathbb{Z}_n^2)^*) \leq C_2' n^2 (\log n)^2. \tag{19.17}$$

More generally, for any dimension $d \geq 3$, there are constants c_d, C_d, c_d' and C_d' such that on $(\mathbb{Z}_n^d)^*$, the relaxation time satisfies

$$c_d n^d \leq t_{\text{rel}}((\mathbb{Z}_n^d)^*) \leq C_d n^d \tag{19.18}$$

and the total variation mixing time satisfies

$$c_d' n^d \log n \leq t_{\text{mix}}(\varepsilon, (\mathbb{Z}_n^d)^*) \leq C_d' n^d \log n. \tag{19.19}$$

PROOF. These follow immediately from combining the bounds on the hitting time and the cover time for tori from Proposition 10.13 and Section 11.3.2, respectively, with Theorems 19.1 and 19.2. ∎

Notes

The results of this chapter are primarily taken from Peres and Revelle (2004), which derives sharper versions of the bounds we discuss, especially in the case of the two-dimensional torus, and also considers the time required for convergence in the uniform metric. The extension of the lower bound on mixing time in Theorem 19.2 to general (rather than vertex-transitive) graphs is new.

Random walks on (infinite) lamplighter groups were analyzed by Kaĭmanovich and Vershik (1983). Their ideas motivate some of the analysis in this chapter.

Scarabotti and Tolli (2008) study the eigenvalues of lamplighter walks. They compute the spectra for the complete graph and the cycle, and use representations of wreath products to give more general results.

Peres and Revelle (2004) also bound the ℓ^∞ mixing time. These bounds were sharpened by Ganapathy and Tetali (2006).

Aldous and Fill (1999, Chapter 4) and Lovász and Winkler (1998) both survey inequalities between Markov chain parameters.

Complements. Recall the discussion in Section 18.4 of cutoff in separation distance.

THEOREM 19.7. *Let (G_n) be a sequence of graphs with $|V_n| = n$. If $t_{\text{hit}}^{(n)} = o\left(t_{\text{cov}}^{(n)}\right)$ as $n \to \infty$, then (G_n^*) has a separation cutoff at time $t_{\text{cov}}^{(n)}$.*

Note that by Theorems 19.1 and 19.2, the hypothesis above implies that $t_{\text{rel}}(G_n^*) = o(t_{\text{mix}}(G_n^*))$.

To prove Theorem 19.7, we will need the following result of Aldous (1991b) on the concentration of the cover time.

THEOREM 19.8 (Aldous). *Let (G_n) be a family of graphs with $|V_n| \to \infty$. If $t_{\text{hit}}^{(n)} = o\left(t_{\text{cov}}^{(n)}\right)$ as $n \to \infty$, then*

$$\frac{\tau_{\text{cov}}^{(n)}}{t_{\text{cov}}^{(n)}} \to 1 \quad \textit{in probablity.}$$

PROOF OF THEOREM 19.7. *Lower bound.* Fix $\varepsilon > 0$ and a starting vertex w. Take $t < (1-\varepsilon)t_{\text{cov}}^{(n)}(G_n)$. Applying Lemma 19.4 to G_n gives

$$s^*(t) \geq \mathbf{P}_w\{\tau_{\text{cov}}^{(n)} > t\} = 1 - \mathbf{P}_w\{\tau_{\text{cov}}^{(n)} \leq t\}.$$

However, Theorem 19.8 implies that $\mathbf{P}_w\{\tau_{\text{cov}}^{(n)} \leq t\}$ goes to 0, so we are done.

Upper bound. Again fix $\varepsilon > 0$, and take $t > (1+2\varepsilon)t_{\text{cov}}^{(n)}$. Then for any vertices v, w and any lamp configuration f we have

$$\Upsilon^t((\mathbf{0}, w), (f, v)) \geq \mathbf{P}_w\{\tau_{\text{cov}}^{(n)} < (1+\varepsilon)t_{\text{cov}}^{(n)}\}2^{-n} \min_{u \in V_n} P^{\varepsilon t_{\text{cov}}^{(n)}}(u, v), \tag{19.20}$$

by conditioning on the location of the lamplighter at time $t - \varepsilon t_{\text{cov}}^{(n)}$ and recalling that once all vertices have been visited, the lamp configuration is uniform.

Theorem 19.8 implies

$$\mathbf{P}_w\{\tau_{\text{cov}}^{(n)} < (1+\varepsilon)t_{\text{cov}}^{(n)}\} = 1 - o(1). \tag{19.21}$$

Theorem 10.14 implies that $t_{\text{mix}} < 3t_{\text{hit}}$ for sufficiently large n, so our initial hypothesis implies that $t_{\text{mix}} = o(\varepsilon t_{\text{cov}}^{(n)})$. Applying Lemma 19.3 now tells us that

$$\min_{u \in V_n} P^{\varepsilon t_{\text{cov}}^{(n)}}(u, v) = \pi(v)(1 - o(1)). \tag{19.22}$$

Taken together (19.20), (19.21), and (19.22) guarantee that the separation distance for the lamplighter chain at time t is $o(1)$. ■

CHAPTER 20

Continuous-Time Chains*

20.1. Definitions

We now construct, given a transition matrix P, a process $(X_t)_{t\in[0,\infty)}$ which we call the ***continuous-time chain*** with transition matrix P. The random times between transitions for this process are i.i.d. exponential random variables of unit rate, and at these transition times moves are made according to P. Continuous-time chains are often natural models in applications, since they do not require transitions to occur at regularly specified intervals.

More precisely, let $T_1, T_2, \ldots$ be independent and identically distributed exponential random variables of unit rate. That is, each T_i takes values in $[0,\infty)$ and has distribution function

$$\mathbf{P}\{T_i \le t\} = \begin{cases} 1 - e^{-t} & \text{if } t \ge 0, \\ 0 & \text{if } t < 0. \end{cases}$$

Let $(\Phi_k)_{k=0}^{\infty}$ be a Markov chain with transition matrix P, independent of the random variables $(T_k)_{k=1}^{\infty}$. Let $S_0 = 0$ and $S_k := \sum_{i=1}^{k} T_i$ for $k \ge 1$. Define

$$X_t := \Phi_k \quad \text{for } S_k \le t < S_{k+1}. \tag{20.1}$$

Change-of-states occur only at the ***transition times*** $S_1, S_2, \ldots$. (Note, however, that if $P(x,x) \ge 0$ for at least one state $x \in \Omega$, then it is possible that the chain does not change state at a transition time.)

Define $N_t := \max\{k : S_k \le t\}$ to be the number of transition times up to and including time t. Observe that $N_t = k$ if and only if $S_k \le t < S_{k+1}$. From the definition (20.1),

$$\mathbf{P}_x\{X_t = y \mid N_t = k\} = \mathbf{P}_x\{\Phi_k = y\} = P^k(x,y). \tag{20.2}$$

Also, the distribution of N_t is Poisson with mean t (Exercise 20.1):

$$\mathbf{P}\{N_t = k\} = \frac{e^{-t}t^k}{k!}. \tag{20.3}$$

The ***heat kernel*** H_t is defined by $H_t(x,y) := \mathbf{P}_x\{X_t = y\}$. From (20.2) and (20.3), it follows that

$$H_t(x,y) = \sum_{k=0}^{\infty} \mathbf{P}_x\{X_t = y \mid N_t = k\}\mathbf{P}_x\{N_t = k\} \tag{20.4}$$

$$= \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} P^k(x,y). \tag{20.5}$$

For an $m \times m$ matrix M, define the $m \times m$ matrix $e^M := \sum_{i=0}^{\infty} \frac{M^i}{i!}$. In matrix representation,

$$H_t = e^{t(P-I)}. \tag{20.6}$$

20.2. Continuous-Time Mixing

The heat kernel for a continuous-time chains like powers of a transition matrix (Theorem 4.9), converges to an equilibrium distribution as $t \to \infty$.

THEOREM 20.1. *Let P be an irreducible transition matrix, and let H_t be the corresponding heat kernel. Then there exists a unique probability distribution π such that $\pi H_t = \pi$ for all $t \geq 0$ and*

$$\max_{x \in \Omega} \|H_t(x,\cdot) - \pi\|_{\text{TV}} \to 0 \quad \textit{as} \quad t \to \infty.$$

REMARK 20.2. Note that the above theorem does not require that P is aperiodic, unlike Theorem 4.9. This is one advantage of working with continuous-time chains.

In view of Theorem 20.1, we define

$$t_{\text{mix}}^{\text{cont}}(\varepsilon) := \inf\left\{t \geq 0 \,:\, \max_{x \in \Omega} \|H_t(x,\cdot) - \pi\|_{TV} \leq \varepsilon\right\}. \tag{20.7}$$

The next theorem, which will imply Theorem 20.1, relates the mixing time of lazy Markov chains with the mixing time of the related continuous-time Markov chain.

THEOREM 20.3. *Let P be an irreducible transition matrix, not necessarily aperiodic or reversible. Let $\widetilde{P} = (1/2)(I + P)$ be the lazy version of P, and let H_t be the heat kernel associated to P. Fix $\varepsilon > 0$.*

(i) *For sufficiently large k, $\|\widetilde{P}^k(x,\cdot) - \pi\|_{\text{TV}} < \varepsilon$ implies $\|H_k(x,\cdot) - \pi\|_{\text{TV}} < 2\varepsilon$.*
(ii) *For sufficiently large m, $\|H_m(x,\cdot) - \pi\|_{\text{TV}} < \varepsilon$ implies $\|\widetilde{P}^{4m}(x,\cdot) - \pi\|_{\text{TV}} < 2\varepsilon$.*

The proof of (ii) in the above theorem requires the following lemma:

LEMMA 20.4. *Let Y be a binomial$(4m, \frac{1}{2})$ random variable, and let $\Psi = \Psi_m$ be a Poisson variable with mean m. Then*

$$\eta_m := \|\mathbf{P}\{Y \in \cdot\} - \mathbf{P}\{\Psi + m \in \cdot\}\|_{\text{TV}} \to 0$$

as $m \to \infty$.

PROOF OF LEMMA 20.4. Note that Y and $\Psi + m$ both have mean $2m$ and variance m. Given $\varepsilon > 0$, let $A = 2\varepsilon^{-1/2}$. By Chebyshev's inequality

$$\mathbf{P}\left\{|Y - 2m| \geq A\sqrt{m}\right\} \leq \varepsilon/4 \quad \text{and} \quad \mathbf{P}\left\{|\Psi - m| \geq A\sqrt{m}\right\} \leq \varepsilon/4. \tag{20.8}$$

Now, using Stirling's formula and computing directly, we can show that uniformly for $|j| \leq A\sqrt{m}$,

$$\mathbf{P}\{Y = 2m + j\} \sim \frac{1}{\sqrt{2\pi m}} e^{-j^2/2m},$$

$$\mathbf{P}\{\Psi + m = 2m + j\} \sim \frac{1}{\sqrt{2\pi m}} e^{-j^2/2m}.$$

Here we write $a_m \sim b_m$ to mean that the ratio a_m/b_m tends to 1 as $m \to \infty$, uniformly for all j such that $|j| \leq A\sqrt{m}$. This follows from the local Central Limit

Theorem (see, for example, Durrett (2005)); or just use Stirling's formula (A.9) — Exercise 20.4 asks for the details.

Thus for large m we have

$$\sum_{|j|\le A\sqrt{m}} \left[\mathbf{P}\{Y = 2m+j\} - \mathbf{P}\{\Psi + m = 2m+j\}\right]$$
$$\le \sum_{|j|\le A\sqrt{m}} \varepsilon \mathbf{P}\{Y = 2m+j\} \le \varepsilon.$$

Dividing this by 2 and using (20.8) establishes the lemma. ■

Proof of Theorem 20.3. (i), *Step 1.* First we show that shortly after the original chain is close to equilibrium, so is the continuous-time chain. Suppose that k satisfies $\|P^k(x,\cdot) - \pi\|_{\mathrm{TV}} < \varepsilon$. Then for arbitrarily small $\delta > 0$ and $t \ge k(1+\delta)$, conditioning on the value of N_t and applying the triangle inequality give

$$\|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} \le \sum_{j\ge 0} \mathbf{P}\{N_t = j\}\, \|P^j(x,\cdot) - \pi\|_{\mathrm{TV}} \le \mathbf{P}\{N_t < k\} + \varepsilon,$$

where the right-hand inequality used monotonicity of $\|P^j(x,\cdot)-\pi\|_{\mathrm{TV}}$ in j. By the Law of large Numbers, $\mathbf{P}\{N_{t(k)} < k\} \to 0$ as $k \to \infty$ for $t(k) \ge k(1+\delta)$. Thus if k is sufficiently large, then $\|H_{t(k)}(x,\cdot) - \pi\|_{\mathrm{TV}} < 2\varepsilon$ for such $t(k)$.

Step 2. Let $\widetilde{H}_t$ be the continuous-time version of the lazy chain $\widetilde{P}$. We claim that $\widetilde{H}_t = H_{t/2}$. There are several ways to see this. One is to observe that H_t involves Ψ_t steps of the lazy chain $\widetilde{P}$. Each of these steps is a step of P with probability 1/2 and a delay with probability 1/2; thinning a Poisson process of rate 1 this way yields a Poisson process of rate 1/2.

Alternatively, the matrix exponentiation of (20.6) yields a very short proof of the claim:

$$\widetilde{H}_t = e^{t(\widetilde{P}-I)} = e^{t(\frac{P+I}{2}-I)} = e^{\frac{t}{2}(P-I)}.$$

Step 3. Now suppose that the lazy chain is close to equilibrium after k steps, that is, $\|\widetilde{P}^k(x,\cdot) - \pi\|_{\mathrm{TV}} < \varepsilon$. We then claim that the continuous-time chain is close to equilibrium shortly after time $k/2$. This is an easy corollary of Steps 1 and 2. If k is large enough, then for $t = \frac{k}{2}(1+\delta)$, we have

$$\|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} = \|\widetilde{H}_{2t} - \pi\|_{\mathrm{TV}} < 2\varepsilon.$$

(ii). Suppose that $\|H_m(x,\cdot) - \pi\|_{\mathrm{TV}} < \varepsilon$; we claim that for sufficiently large m we have $\|\widetilde{P}^{4m}(x,\cdot) - \pi\|_{\mathrm{TV}} < 2\varepsilon$.

After the discrete-time chain has been run for N_m steps, running it for another m steps will not increase the distance to π, so $\|H_m P^m(x,\cdot) - \pi\|_{\mathrm{TV}} < \varepsilon$. (Observe that the matrices H_m and P^m commute.) Now

$$H_m P^m = \sum_{k\ge 0} \mathbf{P}\{\Psi + m = k\} P^k,$$
$$\widetilde{P}^{4m} = \sum_{k\ge 0} \mathbf{P}\{Y = k\} P^k,$$

where Ψ is Poisson(m) and Y is binomial($4m, \frac{1}{2}$). Hence Lemma 20.4 and the coupling description of total variation (Proposition 4.7) give

$$\|H_m P^m(x,\cdot) - \widetilde{P}^{4m}(x,\cdot)\|_{\mathrm{TV}} \le \eta_m,$$

whence

$$\begin{aligned}\|\widetilde{P}^{4m}(x,\cdot) - \pi\|_{\text{TV}} &\leq \|H_m P^m(x,\cdot) - \pi\|_{\text{TV}} + \eta_m \\ &\leq \varepsilon + \eta_m,\end{aligned}$$

as needed. ∎

20.3. Spectral Gap

Given $f \in \mathbb{R}^\Omega$, the function $H_t f : \Omega \to \mathbb{R}$ is defined by

$$(H_t f)(x) := \sum_y H_t(x,y) f(y).$$

The following is a continuous-time version of the inequality (12.8).

LEMMA 20.5. *Let P be a reversible and irreducible transition matrix with spectral gap $\gamma = 1 - \lambda_2$. For $f \in \mathbb{R}^\Omega$,*

$$\|H_t f - E_\pi(f)\|_2^2 \leq e^{-2\gamma t} \operatorname{Var}_\pi(f).$$

PROOF. First, assume that $E_\pi(f) = 0$. One can check directly from (20.5) that

$$\frac{d}{dt} H_t(x,y) = \sum_{z \in \Omega} P(x,z) H_t(z,y) - H_t(x,y),$$

from which it follows that

$$\frac{d}{dt} H_t f(x) = (P - I)(H_t f)(x), \tag{20.9}$$

as anticipated by the identity (20.6). Letting $u(t) := \|H_t f\|_2^2$, from (20.9) it follows that

$$\begin{aligned}u'(t) &= -2 \sum_{x \in \Omega} H_t f(x) \cdot (P - I)(H_t f)(x) \cdot \pi(x) \\ &= -2\langle H_t f, (P - I)(H_t f)\rangle_\pi \\ &= -2\mathcal{E}(H_t f).\end{aligned}$$

Lemma 13.12 implies that $-2\mathcal{E}(H_t f) \leq -2\gamma \|H_t f\|_2^2 = -2u(t)$, whence $u'(t) \leq -2u(t)$. Since $u(0) = \|f\|_2^2$, we conclude that

$$\|H_t f\|_2^2 = u(t) \leq \|f\|_2^2 e^{-2\gamma t}.$$

If $E_\pi(f) \neq 0$, apply the above result to the function $f - E_\pi(f)$. ∎

The following is the continuous-time version of Theorem 12.3.

THEOREM 20.6. *Let P be an irreducible transition matrix with spectral gap γ. Then*

$$|H_t(x,y) - \pi(y)| \leq \sqrt{\frac{\pi(y)}{\pi(x)}} e^{-\gamma t}, \tag{20.10}$$

and so

$$t_{\text{mix}}^{\text{cont}}(\varepsilon) \leq \log\left(\frac{1}{\varepsilon \pi_{\min}}\right) \frac{1}{\gamma}. \tag{20.11}$$

PROOF. If $f_x(y) = \mathbf{1}_{\{y=x\}}/\pi(x)$, then $H_t f_x(y) = H_t(y,x)/\pi(x)$. The reader should check that $\pi(x)H_t(x,y) = \pi(y)H_t(y,x)$, and so $H_t f_x(y) = H_t f_y(x)$. From Lemma 20.5, since $E_\pi(f_x) = 1$ and $\mathrm{Var}_\pi(f_x) = (1-\pi(x))/\pi(x)$, we have

$$\|H_t f_x - 1\|_2^2 \le e^{-2\gamma t}\,\mathrm{Var}_\pi(f) = e^{-2\gamma t}\left(\frac{1-\pi(x)}{\pi(x)}\right) \le \frac{e^{-2\gamma t}}{\pi(x)}. \tag{20.12}$$

Note that

$$\begin{aligned} H_t f_x(y) = \frac{H_t(x,y)}{\pi(y)} &= \frac{\sum_{z\in\Omega} H_{t/2}(x,z)H_{t/2}(z,y)}{\pi(y)} \\ &= \sum_{z\in\Omega} H_{t/2}f_x(z)\cdot H_{t/2}f_z(y)\cdot\pi(z) = \sum_{z\in\Omega} H_{t/2}f_x(z)\cdot H_{t/2}f_y(z)\cdot\pi(z). \end{aligned}$$

Therefore, by Cauchy-Schwarz,

$$\begin{aligned} |H_t f_x(y) - 1| &= \left|\sum_{z\in\Omega}\Big[H_{t/2}f_x(z) - 1\Big]\Big[H_{t/2}f_y(z) - 1\Big]\pi(z)\right| \\ &\le \|H_{t/2}f_x - 1\|_2\,\|H_{t/2}f_y - 1\|_2. \end{aligned}$$

The above with (20.12) shows that

$$\left|\frac{H_t(x,y)}{\pi(y)} - 1\right| \le \frac{e^{-\gamma t}}{\sqrt{\pi(x)\pi(y)}}.$$

Multiplying by $\pi(y)$ gives (20.10)

Summing over y gives

$$2\,\|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} \le e^{-\gamma t}\sum_{y\in\Omega}\frac{\pi(y)}{\sqrt{\pi(y)\pi(x)}} \le \frac{e^{-\gamma t}}{\pi_{\min}}, \tag{20.13}$$

from which follows (20.11) ∎

20.4. Product Chains

For each $i = 1,\dots,n$, let P_i be a reversible transition matrix on Ω_i with stationary distribution $\pi^{(i)}$. Define $\tilde{P}_i$ to be the lift of P_i to $\Omega = \prod_{i=1}^n \Omega_i$: for $\boldsymbol{x} = (x^{(1)},\dots,x^{(n)}) \in \Omega$ and $\boldsymbol{y} = (y^{(1)},\dots,y^{(n)}) \in \Omega$,

$$\tilde{P}_i(\boldsymbol{x},\boldsymbol{y}) := \begin{cases} P_i(x^{(i)},y^{(i)}) & \text{if } x^{(j)} = y^{(j)} \text{ for } j \ne i, \\ 0 & \text{otherwise.} \end{cases} \tag{20.14}$$

We consider the continuous-time chain with transition matrix $P := n^{-1}\sum_{i=1}^n \tilde{P}_i$.

The following gives good upper and lower bounds on $t_{\mathrm{mix}}(\varepsilon)$ for this product chain.

THEOREM 20.7. *Suppose, for $i = 1,\dots,n$, the spectral gap γ_i for the chain with reversible transition matrix P_i is bounded below by γ and the stationary distribution $\pi^{(i)}$ satisfies $\sqrt{\pi^{(i)}_{\min}} \ge c_0$, for some constant $c_0 > 0$. If $P := n^{-1}\sum_{i=1}^n \tilde{P}_i$, where $\tilde{P}_i$ is the matrix defined in (20.14), then the Markov chain with matrix P satisfies*

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\varepsilon) \le \frac{1}{2\gamma}n\log n + \frac{1}{\gamma}n\log(1/[c_0\varepsilon]). \tag{20.15}$$

If the spectral gap $\gamma_i = \gamma$ *for all* i, *then*

$$t_{\text{mix}}^{\text{cont}}(\varepsilon) \geq \frac{n}{2\gamma}\left\{\log n - \log\left[8\log(1/(1-\varepsilon))\right]\right\}. \tag{20.16}$$

COROLLARY 20.8. *For a reversible transition matrix* P *with spectral gap* γ, *let* $P_{(n)} := \frac{1}{n}\sum_{i=1}^{n}\tilde{P}_i$, *where* $\tilde{P}_i$ *is the transition matrix on* Ω^n *defined by*

$$\tilde{P}_i(\boldsymbol{x}, \boldsymbol{y}) = P(x^{(i)}, y^{(i)})\mathbf{1}_{\{x^{(j)}=y^{(j)},\, j\neq i\}}.$$

The family of Markov chains with transition matrices $P_{(n)}$ *has a cutoff at* $\frac{1}{2\gamma}n\log n$.

To obtain a good upper bound on $d(t)$ for product chains, we need to use a distance which is better suited for product distribution than is the total variation distance. For two distributions μ and ν on Ω, define the ***Hellinger affinity*** as

$$I(\mu,\nu) := \sum_{x\in\Omega}\sqrt{\nu(x)\mu(x)}. \tag{20.17}$$

The ***Hellinger distance*** is defined as

$$d_H(\mu,\nu) := \sqrt{2 - 2I(\mu,\nu)}. \tag{20.18}$$

Note also that

$$d_H(\mu,\nu) = \sqrt{\sum_{x\in\Omega}\left(\sqrt{\mu(x)} - \sqrt{\nu(x)}\right)^2}. \tag{20.19}$$

The measure ν ***dominates*** μ if $\nu(x) = 0$ implies $\mu(x) = 0$, in which case we write $\mu \ll \nu$. If $\mu \ll \nu$, then we can define $g(x) := \frac{\mu(x)}{\nu(x)}\mathbf{1}_{\{\nu(x)>0\}}$, and we also have the identity

$$d_H(\mu,\nu) = \|\sqrt{g} - 1\|_{\ell^2(\nu)}. \tag{20.20}$$

The following lemma shows why the Hellinger distance is useful for product measure.

LEMMA 20.9. *For measures* $\mu^{(i)}$ *and* $\nu^{(i)}$ *on* Ω_i, *let* $\mu := \prod_{i=1}^{n}\mu^{(i)}$ *and* $\nu := \prod_{i=1}^{n}\nu^{(i)}$. *The Hellinger affinity satisfies*

$$I(\mu,\nu) = \prod_{i=1}^{n} I(\mu^{(i)},\nu^{(i)}),$$

and therefore

$$d_H^2(\mu,\nu) \leq \sum_{i=1}^{n} d_H^2(\mu^{(i)},\nu^{(i)}). \tag{20.21}$$

The proof is left as Exercise 20.5.

We will also need to compare Hellinger with other distances.

LEMMA 20.10. *Let* μ *and* ν *be probability distributions on* Ω. *The total variation distance and Hellinger distance satisfy*

$$\|\mu - \nu\|_{\text{TV}} \leq d_H(\mu,\nu). \tag{20.22}$$

If $\mu \ll \nu$, *then*

$$d_H(\mu,\nu) \leq \|g - 1\|_{\ell^2(\nu)}, \tag{20.23}$$

where $g(x) = \frac{\mu(x)}{\nu(x)}\mathbf{1}_{\{\mu(x)>0\}}$.

PROOF. First, observe that

$$\|\mu - \mu\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)|$$
$$= \frac{1}{2}\sum_{x\in\Omega} |\sqrt{\mu(x)} - \sqrt{\nu(x)}| \left(\sqrt{\mu(x)} + \sqrt{\nu(x)}\right). \tag{20.24}$$

By the Cauchy-Schwarz inequality,

$$\sum_{x\in\Omega}\left(\sqrt{\mu(x)} + \sqrt{\nu(x)}\right)^2 = 2 + 2\sum_{x\in\Omega}\sqrt{\mu(x)\nu(x)} \leq 4. \tag{20.25}$$

Applying Cauchy-Schwarz on the right-hand side of (20.24) and using the bound (20.25) shows that

$$\|\mu - \nu\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left[\sum_{x\in\Omega}\left(\sqrt{\mu(x)} - \sqrt{\nu(x)}\right)^2\right] 4 = d_H^2(\mu,\nu).$$

To prove (20.23), use (20.20) and the inequality $(1-\sqrt{x})^2 \leq (1-x)^2$, valid for all x:

$$d_H(\mu,\nu) = \|\sqrt{g} - 1\|_2 \leq \|g - 1\|_2.$$

■

We will also make use of the following lemma, useful for obtaining lower bounds. This is the continuous-time version of the bound (12.13) in the proof of Theorem 12.3.

LEMMA 20.11. *Let P be an irreducible reversible transition matrix, and let H_t be the heat kernel of the associated continuous-time Markov chain. If λ is an eigenvalue of P, then*

$$\max_{x\in\Omega} \|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} \geq \frac{1}{2}e^{-(1-\lambda)t}. \tag{20.26}$$

PROOF. Let f be an eigenfunction of P with eigenvalue λ. We have that

$$H_t f(x) = \sum_{k=0}^{\infty} e^{-t}\frac{t^k}{k!}P^k f(x) = e^{-t}\sum_{k=0}^{\infty}\frac{(t\lambda)^k}{k!}f(x) = e^{-t(1-\lambda)}f(x).$$

Since f is orthogonal to $\mathbf{1}$, we have $\sum_{y\in\Omega} f(y)\pi(y) = 0$, whence

$$e^{-t(1-\lambda)}|f(x)| = |H_t f(x)|$$
$$= \left|\sum_{y\in\Omega}[H_t(x,y)f(y) - \pi(y)f(y)]\right|$$
$$\leq \|f\|_\infty 2\,\|H_t(x,\cdot) - \pi\|_{\mathrm{TV}}.$$

Taking x with $f(x) = \|f\|_\infty$ yields (20.26). ■

PROOF OF THEOREM 20.7. *Proof of* (20.15). Let $\boldsymbol{X}_t = (X_t^{(1)}, \ldots, X_t^{(n)})$ be the Markov chain with transition matrix P and heat kernel H_t. Note that

$$H_t = \prod_{i=1}^{n} e^{(t/n)(\tilde{P}_i - I)},$$

which follows from Exercise 20.3 since $\tilde{P}_i$ and $\tilde{P}_j$ commute. Therefore, for $\boldsymbol{x}, \boldsymbol{y} \in \Omega$,

$$\mathbf{P}_{\boldsymbol{x}}\{\boldsymbol{X}_t = \boldsymbol{y}\} = H_t(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} e^{(t/n)(\tilde{\mathbf{P}}_i - I)}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} \mathbf{P}_{\boldsymbol{x}}\{X^{(i)}_{t/n} = y^{(i)}\}. \quad (20.27)$$

Since (20.27) implies that $H_t(\boldsymbol{x}, \cdot) = \prod_{i=1}^{n} H^{(i)}_{t/n}(x^{(i)}, \cdot)$, by (20.21),

$$d_H^2(H_t(\boldsymbol{x}, \cdot), \boldsymbol{\pi}) \le \sum_{i=1}^{n} d_H^2\big(H^{(i)}_{t/n}(x^{(i)}, \cdot),\ \pi^{(i)}\big).$$

Using (20.22) and (20.23) together with the above inequality shows that

$$\|H_t(\boldsymbol{x}, \cdot) - \boldsymbol{\pi}\|^2_{\mathrm{TV}} \le \sum_{i=1}^{n} \left\| \frac{H^{(i)}_t(x^{(i)}, \cdot)}{\pi^{(i)}} - 1 \right\|_2^2 .$$

Combining the above with (20.12) and using the hypotheses of the theorem yields

$$\|H_t(\boldsymbol{x}, \cdot) - \boldsymbol{\pi}\|^2_{\mathrm{TV}} \le \sum_{i=1}^{n} \frac{e^{-2\gamma_i t}}{\pi^{(i)}(x^{(i)})} \le \frac{n e^{-2\gamma t}}{c_0^2}.$$

In particular,

$$\|H_t(\boldsymbol{x}, \cdot) - \boldsymbol{\pi}\|_{\mathrm{TV}} \le \frac{\sqrt{n} e^{-\gamma t}}{c_0},$$

from which follows (20.15).

Proof of (20.16). Pick $x_0^{(i)}$ which maximizes $\left\| H^{(i)}_t(x, \cdot) - \pi^{(i)} \right\|_{\mathrm{TV}}$. From (20.22), it follows that

$$I\left(H^{(i)}_{t/n}(x_0^{(i)}, \cdot),\ \pi^{(i)}\right) \le 1 - \frac{1}{2}\left\| H^{(i)}_{t/n}(x_0^{(i)}, \cdot) - \pi^{(i)} \right\|^2_{\mathrm{TV}} .$$

Applying Lemma 20.11 and using the above inequality shows that

$$I\left(H^{(i)}_{t/n}(x_0^{(i)}, \cdot),\ \pi^{(i)}\right) \le 1 - \frac{e^{-2\gamma t/n}}{8}.$$

Let $\boldsymbol{x}_0 := \big(x_0^{(1)}, \ldots, x_0^{(n)}\big)$. By Lemma 20.9,

$$I\left(H_t(\boldsymbol{x}_0, \cdot), \boldsymbol{\pi}\right) \le \left(1 - \frac{e^{-2\gamma t/n}}{8}\right)^n . \quad (20.28)$$

Note that by (4.13), for any two distributions μ and ν,

$$I(\mu, \nu) = \sum_{x \in \Omega} \sqrt{\mu(x)\nu(x)} \ge \sum_{x \in \Omega} \mu(x) \wedge \nu(x) = 1 - \|\mu - \nu\|_{\mathrm{TV}} ,$$

and consequently,

$$\|\mu - \nu\|_{\mathrm{TV}} \ge 1 - I(\mu, \nu). \quad (20.29)$$

Using (20.29) in (20.28) shows that

$$\|H_t(\boldsymbol{x}_0, \cdot) - \boldsymbol{\pi}\|_{\mathrm{TV}} \ge 1 - \left(1 - \frac{e^{-2\gamma t/n}}{8}\right)^n .$$

Therefore, if

$$t < \frac{n}{2\gamma}\left\{\log n - \log\left[8 \log(1/(1-\varepsilon))\right]\right\},$$

then

$$\|H_t(\boldsymbol{x}_0, \cdot) - \boldsymbol{\pi}\|_{\mathrm{TV}} > \varepsilon.$$

That is, (20.16) holds. ∎

Exercises

EXERCISE 20.1. Let $T_1, T_2, \dots$ be an i.i.d. sequence of exponential random variables of rate μ, let $S_k = \sum_{i=1}^k T_i$, and let $N_t = \max\{k : S_k \le t\}$.

(a) Show that S_k has a gamma distribution with shape parameter k and rate parameter μ, i.e. its density function is

$$f_k(s) = \frac{\mu^k s^{k-1} e^{-\mu s}}{(k-1)!}.$$

(b) Show by computing $\mathbf{P}\{S_k \le t < S_{k+1}\}$ that N_t is a Poisson random variable with mean μt.

EXERCISE 20.2. We outline below an alternative proof that N_t has a Poisson distribution with mean t; fill in the details.

Divide the interval $[0, t]$ into t/Δ subintervals of length Δ. The chance of at least one transition in each subinterval is

$$1 - e^{-t/\Delta} = t/\Delta + O((t/\Delta)^2),$$

and the chance of more than one transition is $O((t/\Delta)^2)$. The number of transitions recorded in subintervals are independent of one another. Therefore, as $\Delta \to 0$, the total number of arrivals tends to a Poisson distribution with parameter t.

EXERCISE 20.3. Show that if A and B are $m \times m$ matrices which commute, then $e^{A+B} = e^A e^B$.

EXERCISE 20.4. Let Y be a binomial random variable with parameters $4m$ and $1/2$. Show that

$$\mathbf{P}\{Y = 2m + j\} = \frac{1}{\sqrt{2\pi m}} e^{-j^2/2n} [1 + \varepsilon_m],$$

where $\varepsilon_m \to 0$ uniformly for $j/\sqrt{m} \le A$.

EXERCISE 20.5. Show that if $\mu = \prod_{i=1}^n \mu_i$ and $\nu = \prod_{i=1}^n \nu_i$, then

$$I(\mu, \nu) = \prod_{i=1}^n I(\mu_i, \nu_i),$$

and therefore

$$d_H^2(\mu, \nu) \le \sum_{i=1}^n d_H^2(\mu_i, \nu_i).$$

Notes

To make the estimates in Section 20.2 more quantitative, one needs an estimate of the convergence rate for η_m in Lemma 20.4. This can be done in at least three ways:

(1) We could apply a version of Stirling's formula with error bounds (see (A.10)) in conjunction with large deviation estimates for Y and Ψ.
(2) We could replace Stirling's formula with a precise version of the local Central Limit Theorem; see e.g. Spitzer (1976).

(3) One can also use Stein's method; see Chyakanavichyus and Vaĭtkus (2001) or Röllin (2007).

These methods all show that η_m is of order $m^{-1/2}$.

Mixing of product chains is studied in detail in Barrera, Lachaud, and Ycart (2006). The Hellinger distance was used by Kakutani (1948) to characterize when two product measures on an infinite product space are singular.

CHAPTER 21

Countable State Space Chains*

In this chapter we treat the case where Ω is not necessarily finite, although we assume it is a countable set. A classical example is the simple random walk on $\mathbb{Z}^d$, which we have already met in the case $d = 1$ in Section 2.7. This walk moves on $\mathbb{Z}^d$ by choosing uniformly at random among her $2d$ nearest neighbors. There is a striking dependence on the dimension d: when $d \geq 3$, the walk may wander off "to infinity", never returning to its starting place, while this is impossible in dimensions $d \leq 2$. We will return to this example later.

As before, P is a function from $\Omega \times \Omega$ to $[0, 1]$ satisfying $\sum_{y \in \Omega} P(x, y) = 1$ for all $x \in \Omega$. We still think of P as a matrix, except now it has countably many rows and columns. The matrix arithmetic in the finite case extends to the countable case without any problem, as do the concepts of irreducibility and aperiodicity. The joint distribution of the infinite sequence (X_t) is still specified by P along with a starting distribution μ on Ω.

21.1. Recurrence and Transience

EXAMPLE 21.1 (Simple random walk on $\mathbb{Z}$). Let (X_t) have transition matrix

$$P(j, k) = \begin{cases} 1/2 & \text{if } k = j \pm 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let A_k be the event that the walk started from zero reaches absolute value 2^k before it returns to zero. By symmetry, $\mathbf{P}_0(A_1) = 1/2$ and $\mathbf{P}_0(A_{k+1} \mid A_k) = 1/2$. Thus $\mathbf{P}_0(A_k) = 2^{-k}$, and in particular

$$\mathbf{P}_0\{\tau_0^+ = \infty\} = \mathbf{P}_0\left(\bigcap_{k=1}^{\infty} A_k\right) = \lim_{k \to \infty} \mathbf{P}_0(A_k) = 0.$$

The penultimate equality follows since the events $\{A_k\}$ are decreasing.

EXAMPLE 21.2 (Biased random walk on $\mathbb{Z}$). Suppose now that a particle on $\mathbb{Z}$ makes biased moves, so that

$$P(j, k) = \begin{cases} q & \text{for } k = j - 1, \\ p & \text{for } k = j + 1, \end{cases}$$

where $q < p$ and $q + p = 1$. Recall the gambler's ruin formula (9.21) for biased random walk,

$$\mathbf{P}_k\{\tau_n < \tau_0\} = \frac{1 - (q/p)^k}{1 - (q/p)^n}.$$

Thus,

$$\mathbf{P}_1\{\tau_0 = \infty\} \geq \mathbf{P}_1\left(\bigcap_{n=2}^{\infty}\{\tau_n < \tau_0\}\right) = \lim_n \frac{1-(q/p)}{1-(q/p)^n} = \frac{p-q}{p} > 0.$$

Since $\mathbf{P}_0\{\tau_0 = \infty\} = \mathbf{P}_1\{\tau_0 = \infty\}$, there is a positive probability that the biased random walk never returns to its starting position.

This is also a consequence of the Strong Law of Large Numbers; see Exercise 21.1.

We have seen that on $\mathbb{Z}$ the unbiased random walk (Example 21.1) and the biased random walk (Example 21.2) have quite different behavior. We make the following definition to describe this difference.

We define a state $x \in \Omega$ to be ***recurrent*** if $\mathbf{P}_x\{\tau_x^+ < \infty\} = 1$. Otherwise, x is called ***transient***.

PROPOSITION 21.3. *Suppose that P is the transition matrix of an irreducible Markov chain (X_t). Define $G(x,y) := \mathbf{E}_x\left(\sum_{t=0}^{\infty} \mathbf{1}_{\{X_t=y\}}\right) = \sum_{t=0}^{\infty} P^t(x,y)$ to be the expected number of visits to y starting from x. The following are equivalent:*

(i) $G(x,x) = \infty$ *for some* $x \in \Omega$.
(ii) $G(x,y) = \infty$ *for all* $x, y \in \Omega$.
(iii) $\mathbf{P}_x\{\tau_x^+ < \infty\} = 1$ *for some* $x \in \Omega$.
(iv) $\mathbf{P}_x\{\tau_y^+ < \infty\} = 1$ *for all* $x, y \in \Omega$.

PROOF. Every time the chain visits x, it has the same probability of eventually returning to x, independent of the past. Thus the number of visits to x is a geometric random variable with success probability $1 - \mathbf{P}_x\{\tau_x^+ < \infty\}$. It follows that (i) and (iii) are equivalent.

Suppose $G(x_0, x_0) = \infty$, and let $x, y \in \Omega$. By irreducibility, there exist r and s such that $P^r(x, x_0) > 0$ and $P^s(x_0, y) > 0$. Then

$$\begin{aligned} P^r(x,x_0)P^t(x_0,x_0)P^s(x_0,y) &= \mathbf{P}_x\{X_r = x_0,\ X_{r+t} = x_0,\ X_{r+t+s} = y\} \\ &\leq \mathbf{P}_x\{X_{r+t+s} = y\} = P^{r+t+s}(x,y). \end{aligned}$$

Thus,

$$G(x,y) \geq \sum_{t=0}^{\infty} P^{r+t+s}(x,y) = P^r(x,x_0)P^s(x_0,y)\sum_{t=0}^{\infty} P^t(x_0,x_0). \tag{21.1}$$

Since $P^r(x,x_0)P^s(x_0,y) > 0$, (21.1) shows that conditions (i) and (ii) are equivalent.

Suppose that $\mathbf{P}_{x_0}\{\tau_{x_0}^+ < \infty\} = 1$ for some $x_0 \in \Omega$, and let $x, y \in \Omega$.

If $\mathbf{P}_{x_0}\{\tau_x < \tau_{x_0}^+\} = 0$, then x is never hit when starting from x_0, contradicting the irreducibility of the chain. We have

$$0 = \mathbf{P}_{x_0}\{\tau_{x_0}^+ = \infty\} \geq \mathbf{P}_{x_0}\{\tau_x < \tau_{x_0}^+\}\mathbf{P}_x\{\tau_{x_0}^+ = \infty\}.$$

Since $\mathbf{P}_{x_0}\{\tau_x < \tau_{x_0}^+\} > 0$, it must be true that $\mathbf{P}_x\{\tau_{x_0}^+ = \infty\} = 0$. Each time the chain visits x_0, it has positive probability of visiting y, independent of the past. Since the chain visits x_0 infinitely often, it will eventually visit y. To summarize: starting from x, the chain is certain to visit x_0, and starting from x_0, the chain is certain to visit y. Consequently, $\mathbf{P}_x\{\tau_y < \infty\} = 1$. We conclude that (iii) and (iv) are equivalent. ■

By Proposition 21.3, for an irreducible chain, a single state is recurrent if and only if all states are recurrent. For this reason, an irreducible chain can be classified as either recurrent or transient.

EXAMPLE 21.4 (Simple random walk on $\mathbb{Z}$ revisited). Another proof that the simple random walker on $\mathbb{Z}$ discussed in Example 21.1 is recurrent uses Proposition 21.3.

When started at 0, the walk can return to 0 only at even times, with the probability of returning after $2t$ steps equal to $\mathbf{P}_0\{X_{2t} = 0\} = \binom{2t}{t}2^{-2t}$. By application of Stirling's formula (A.9), $\mathbf{P}_0\{X_{2t} = 0\} \sim ct^{-1/2}$. Then

$$G(0,0) = \sum_{t=0}^{\infty} \mathbf{P}_0\{X_{2t} = 0\} = \infty,$$

so by Proposition 21.3 the chain is recurrent.

EXAMPLE 21.5. The simple random walk on $\mathbb{Z}^2$ moves at each step by selecting each of the four neighboring locations with equal probability. Instead, consider at first the "corner" walk, which at each move adds with equal probability one of $\{(1,1),(1,-1),(-1,1),(-1,-1)\}$ to the current location. The advantage of this walk is that its coordinates are independent simple random walks on $\mathbb{Z}$. So

$$\mathbf{P}_{(0,0)}\{X_{2t} = (0,0)\} = \mathbf{P}_{(0,0)}\left\{X_{2t}^1 = 0\right\}\mathbf{P}_{(0,0)}\left\{X_{2t}^2 = 0\right\} \sim \frac{c}{n}.$$

Again by Proposition 21.3, the chain is recurrent. Now notice that the usual nearest-neighbor simple random walk is a rotation of the corner walk by $\pi/4$, so it is recurrent.

For random walks on infinite graphs, the electrical network theory of Chapter 9 is very useful for deciding if a chain is recurrent.

21.2. Infinite Networks

For an infinite connected graph $G = (V, E)$ with edge conductances $\{c(e)\}_{e\in E}$, let $a \in V$, and let $\{G_n = (V_n, E_n)\}$ be a sequence of finite connected subgraphs containing a such that

(i) E_n contains all edges in E with both endpoints in V_n,
(ii) $V_n \subset V_{n+1}$ for all n, and
(iii) $\bigcup_{n=1}^{\infty} V_n = V$.

For each n, construct a modified network $G_n^\star$ in which all the vertices in $V \setminus V_n$ are replaced by a single vertex z_n (adjacent to all vertices in V_n which are adjacent to vertices in $V \setminus V_n$), and define

$$\mathcal{R}(a \leftrightarrow \infty) := \lim_{n\to\infty} \mathcal{R}\left(a \leftrightarrow z_n \text{ in } G_n^\star\right).$$

The limit above exists and does not depend on the sequence $\{G_n\}$ by Rayleigh's Monotonicity Principle. Define $\mathcal{C}(a \leftrightarrow \infty) := [\mathcal{R}(a \leftrightarrow \infty)]^{-1}$. By (9.13),

$$\mathbf{P}_a\{\tau_a^+ = \infty\} = \lim_{n\to\infty} \mathbf{P}_a\{\tau_{z_n} < \tau_a^+\} = \lim_{n\to\infty} \frac{\mathcal{C}(a \leftrightarrow z_n)}{\pi(a)} = \frac{\mathcal{C}(a \leftrightarrow \infty)}{\pi(a)}.$$

The first and fourth expressions above refer to the network G, while the second and third refer to the networks $G_n^\star$.

A flow on G from a to infinity is an antisymmetric edge function obeying the node law at all vertices except a. Thomson's Principle (Theorem 9.10) remains valid for infinite networks:

$$\mathcal{R}(a \leftrightarrow \infty) = \inf\left\{\mathcal{E}(\theta) \,:\, \theta \text{ a unit flow from } a \text{ to } \infty\right\}. \tag{21.2}$$

As a consequence, Rayleigh's Monotonicity Law (Theorem 9.12) also holds for infinite networks.

The following summarizes the connections between resistance and recurrence.

PROPOSITION 21.6. *Let $\langle G, \{c(e)\}\rangle$ be a network. The following are equivalent:*

(i) *The weighted random walk on the network is transient.*
(ii) *There is some node a with $\mathcal{C}(a \leftrightarrow \infty) > 0$. (Equivalently, $\mathcal{R}(a \leftrightarrow \infty) < \infty$.)*
(iii) *There is a flow θ from some node a to infinity with $\|\theta\| > 0$ and $\mathcal{E}(\theta) < \infty$.*

In an infinite network $\langle G, \{c_e\}\rangle$, a version of Proposition 9.15 (the Nash-Williams inequality) is valid.

PROPOSITION 21.7 (Nash-Williams). *If there exist disjoint edge-cutsets $\{\Pi_n\}$ that separate a from ∞ and satisfy*

$$\sum_n \left(\sum_{e \in \Pi_n} c(e)\right)^{-1} = \infty, \tag{21.3}$$

then the weighted random walk on $\langle G, \{c_e\}\rangle$ is recurrent.

PROOF. Recall the definition of z_n given in the beginning of this section. The assumption (21.3) implies that $\mathcal{R}(a \leftrightarrow z_n) \to \infty$. Consequently, by Proposition 9.5, $\mathbf{P}_a\{\tau_{z_n} < \tau_a^+\} \to 0$, and the chain is recurrent. ∎

EXAMPLE 21.8 ($\mathbb{Z}^2$ is recurrent). Take $c(e) = 1$ for each edge of $G = \mathbb{Z}^2$ and consider the cutsets consisting of edges joining vertices in $\partial\square_n$ to vertices in $\partial\square_{n+1}$, where $\square_n := [-n, n]^2$. Then by the Nash-Williams inequality,

$$\mathcal{R}(a \leftrightarrow \infty) \geq \sum_n \frac{1}{4(2n+1)} = \infty.$$

Thus, simple random walk on $\mathbb{Z}^2$ is recurrent. Moreover, we obtain a lower bound for the resistance from the center of a square $\square_n = [-n, n]^2$ to its boundary:

$$\mathcal{R}(0 \leftrightarrow \partial\square_n) \geq c \log n.$$

EXAMPLE 21.9 ($\mathbb{Z}^3$ is transient). To each directed edge $\vec{e}$ in the lattice $\mathbb{Z}^3$, attach an orthogonal unit square $\square_e$ intersecting $\vec{e}$ at its midpoint m_e. Define $\theta(\vec{e})$ to be the area of the radial projection of $\square_e$ onto the sphere of radius $1/4$ centered at the origin, taken with a positive sign if $\vec{e}$ points in the same direction as the radial vector from 0 to m_e and with a negative sign otherwise (see Figure 21.1). By considering the projections of all faces of the unit cube centered at a lattice point, we can easily verify that θ satisfies the node law at all vertices except the origin. Hence θ is a flow from 0 to ∞ in $\mathbb{Z}^3$. It is easy to bound its energy:

$$\mathcal{E}(\theta) \leq \sum_n C_1 n^2 \left(\frac{C_2}{n^2}\right)^2 < \infty.$$

By Proposition 21.6, $\mathbb{Z}^3$ is transient. This works for any $\mathbb{Z}^d$, $d \geq 3$. An analytic description of the same flow was given by T. Lyons (1983).

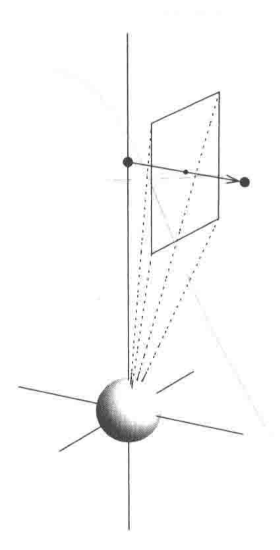


FIGURE 21.1. Projecting a unit square orthogonal to the directed edge $((0,0,2),(1,0,2))$ onto the sphere of radius $1/4$ centered at the origin.

21.3. Positive Recurrence and Convergence

The Convergence Theorem as stated in Theorem 4.9 does not hold for all irreducible and aperiodic chains on infinite state spaces. If the chain is transient, then by Proposition 21.3, $\sum_{t=0}^{\infty} \mathbf{P}_x\{X_t = y\} < \infty$ for all $x, y \in X$. This implies that for all $x, y \in \Omega$,

$$\lim_{t\to\infty} \mathbf{P}_x\{X_t = y\} = 0. \tag{21.4}$$

That is, if there is a probability π on Ω such that $(\mu P^t)(x) \to \pi(x)$ for all $x \in \Omega$, then the chain must be recurrent.

However, recurrence is not sufficient. For example, the simple random walker of Example 21.4, a recurrent chain, also satisfies (21.4). A condition stronger than recurrence is required.

EXAMPLE 21.10. We have already seen that the simple random walker on $\mathbb{Z}$ is recurrent. Let $\alpha = \mathbf{E}_1(\tau_0)$. By conditioning on the first move of the walk,

$$\alpha = \frac{1}{2}[1] + \frac{1}{2}[1 + \mathbf{E}_2(\tau_0)] = 1 + \alpha.$$

The last equality follows since the time to go from 2 to 0 equals the time to go from 2 to 1 plus the time to go from 1 to 0, and the time to go from 2 to 1 has the same distribution as the time to go from 1 to 0. There is no finite number α which satisfies this equation, so we must have $\alpha = \infty$. From this it follows that $\mathbf{E}_0(\tau_0^+) = \infty$. Thus, although τ_0 is a finite random variable with probability one, it has infinite expectation.

A state x is called ***positive recurrent*** if $\mathbf{E}_x(\tau_x^+) < \infty$. As Example 21.10 shows, this property is strictly stronger than recurrence.

PROPOSITION 21.11. *If (X_t) is a Markov chain with irreducible transition matrix P, then the following are equivalent:*

(i) $\mathbf{E}_x(\tau_x^+) < \infty$ *for some* $x \in \Omega$.
(ii) $\mathbf{E}_x(\tau_y^+) < \infty$ *for all* $x, y \in \Omega$.

PROOF. Suppose that $\mathbf{E}_y(\tau_y^+) < \infty$. On the event $\{\tau_x < \tau_y^+\}$, the return time to y satisfies $\tau_y^+ = \tau_x + \tilde{\tau}_y^+$, where $\tilde{\tau}_y^+$ is the amount of time after τ_x^+ until the chain first hits y. The chain from τ_x^+ onwards is just like a chain started from x at time 0. Therefore, given that $\tau_x^+ < \tau_y^+$, the distribution of $\tilde{\tau}_y^+$ is the same as the distribution of τ_y^+ when the chain is started at x. We conclude that

$$\infty > \mathbf{E}_y \tau_y^+ \geq \mathbf{E}_y(\tau_y^+ \mid \tau_x < \tau_y^+)\mathbf{P}_y\{\tau_x < \tau_y^+\} \geq \mathbf{E}_x(\tau_y^+)\mathbf{P}_y\{\tau_x < \tau_y^+\}.$$

By irreducibility, $\mathbf{P}_y\{\tau_x < \tau_y^+\} > 0$, whence $\mathbf{E}_x(\tau_y^+) < \infty$.

Now let x and y be any two states in Ω. Define

$$\tau_{a \to b} = \inf\{t > \tau_a^+ \,:\, X_t = b\},$$

the first time after first visiting a that the chain visits b. Observe that

$$\infty > \mathbf{E}_{x_0}(\tau_{x \to x_0}) = \mathbf{E}_{x_0}(\tau_x^+) + \mathbf{E}_x(\tau_{x_0}^+).$$

Consequently, both $\mathbf{E}_{x_0}(\tau_x^+)$ and $\mathbf{E}_x(\tau_{x_0}^+)$ are finite for any x. It follows that

$$\mathbf{E}_x(\tau_y^+) \leq \mathbf{E}_x(\tau_{x_0}^+) + \mathbf{E}_{x_0}(\tau_y^+) < \infty.$$

∎

Thus if a single state of the chain is positive recurrent, all states are positive recurrent. We can therefore classify an irreducible chain as positive recurrent if one state and hence all states are positive recurrent. A chain which is recurrent but not positive recurrent is called ***null recurrent***.

The following relates positive recurrence to the existence of a stationary distribution:

THEOREM 21.12. *An irreducible Markov chain with transition matrix P is positive recurrent if and only if there exists a probability distribution π on Ω such that $\pi = \pi P$.*

LEMMA 21.13 (Kac). *Let (X_t) be an irreducible Markov chain with transition matrix P. Suppose that there is a stationary distribution π solving $\pi = \pi P$. Then for any set $S \subset \Omega$,*

$$\sum_{x \in S} \pi(x)\mathbf{E}_x(\tau_S^+) = 1. \tag{21.5}$$

In other words, the expected return time to S when starting at the stationary distribution conditioned on S is $\pi(S)^{-1}$.

PROOF. Let (Y_t) be the reversed chain with transition matrix $\hat{P}$, defined in (1.33).

First we show that both (X_t) and (Y_t) are recurrent. Fix a state x and define

$$\alpha(t) := \mathbf{P}_\pi\{X_t = x,\, X_s \neq x \text{ for } s > t\}.$$

By stationarity,

$$\alpha(t) = \mathbf{P}_\pi\{X_t = x\}\mathbf{P}_x\{\tau_x^+ = \infty\} = \pi(x)\mathbf{P}_x\{\tau_x^+ = \infty\}. \tag{21.6}$$

Since the events $\{X_t = x,\ X_s \neq x \text{ for } s > t\}$ are disjoint for distinct t,

$$\sum_{t=0}^{\infty} \alpha(t) \leq 1.$$

Since it is clear from (21.6) that $\alpha(t)$ does not depend on t, it must be that $\alpha(t) = 0$ for all t. From the identity (21.6) and Exercise 21.2, it follows that $\mathbf{P}_x\{\tau_x^+ < \infty\} = 1$. The same argument works for the reversed chain as well, so (Y_t) is also recurrent.

For $x \in S, y \in \Omega$ and $t \geq 0$, sum the identity

$$\pi(z_0)P(z_0, z_1)P(z_1, z_2)\cdots P(z_{t-1}, z_t) = \pi(z_t)\hat{P}(z_t, z_{t-1})\cdots \hat{P}(z_1, z_0)$$

over all sequences where $z_0 = x$, the states $z_1, \ldots, z_{t-1}$ are not in S, and $z_t = y$ to obtain

$$\pi(x)\mathbf{P}_x\{\tau_S^+ \geq t,\ X_t = y\} = \pi(y)\hat{\mathbf{P}}_y\{\tau_S^+ = t,\ Y_t = x\}. \tag{21.7}$$

(We write $\hat{\mathbf{P}}$ for the probability measure corresponding to the reversed chain.) Summing over all $x \in S$, $y \in \Omega$, and $t \geq 0$ shows that

$$\sum_{x \in S} \pi(x) \sum_{t=1}^{\infty} \mathbf{P}_x\{\tau_S^+ \geq t\} = \hat{\mathbf{P}}_\pi\{\tau_S^+ < \infty\} = 1.$$

(The last equality follows from recurrence of (Y_t).) Since τ_S^+ takes only positive integer values, this simplifies to

$$\sum_{x \in S} \pi(x)\mathbf{E}_x\{\tau_S^+\} = 1. \tag{21.8}$$

■

PROOF OF THEOREM 21.12. That the chain is positive recurrent when a stationary distribution exists follows from Lemma 21.13 and Exercise 21.2.

The key fact needed to show that $\tilde{\pi}$ defined in (1.19) can be normalized to yield a stationary distribution is that $\mathbf{E}_z(\tau_z^+) < \infty$, which holds now by positive recurrence. Thus the proof that a stationary distribution exists goes through as in the finite case (Proposition 1.14). ■

THEOREM 21.14. *Let P be an irreducible and aperiodic transition matrix for a Markov chain (X_t). If the chain is positive recurrent, then there is a unique probability distribution π on Ω such that $\pi = \pi P$ and for all $x \in \Omega$,*

$$\lim_{t \to \infty} \|P^t(x, \cdot) - \pi\|_{TV} = 0. \tag{21.9}$$

PROOF. The existence of π solving $\pi = \pi P$ is one direction of Theorem 21.12.

We now show that for any two states x and y we can couple together the chain started from x with the chain started from y so that the two chains eventually meet with probability one.

Consider the chain on $\Omega \times \Omega$ with transition matrix

$$\tilde{P}((x, y), (z, w)) = P(x, z)P(y, w), \quad \text{for all } (x, y) \in \Omega \times \Omega,\ (z, w) \in \Omega \times \Omega. \tag{21.10}$$

This chain makes independent moves in the two coordinates, each according to the matrix P. Aperiodicity implies that this chain is irreducible (see Exercise 21.5). If (X_t, Y_t) is a chain started with product distribution $\mu \times \nu$ and run with transition matrix $\tilde{P}$, then (X_t) is a Markov chain with transition matrix P and initial distribution μ, and (Y_t) is a Markov chain with transition matrix P and initial distribution ν.

Note that

$$(\pi \times \pi)\tilde{P}(z,w) = \sum_{(x,y)\in\Omega\times\Omega} (\pi \times \pi)(x,y)P(x,z)P(y,w)$$
$$= \sum_{x\in\Omega} \pi(x)P(x,z) \sum_{y\in\Omega} \pi(y)P(y,w).$$

Since $\pi = \pi P$, the right-hand side equals $\pi(z)\pi(w) = (\pi \times \pi)(z,w)$. Thus $\pi \times \pi$ is a stationary distribution for $\tilde{P}$. By Theorem 21.12, the chain (X_t, Y_t) is positive recurrent. In particular, for any fixed x_0, if

$$\tau := \min\{t > 0 : (X_t, Y_t) = (x_0, x_0)\},$$

then

$$\mathbf{P}_{x,y}\{\tau < \infty\} = 1 \quad \text{for all } x, y \in \Omega. \tag{21.11}$$

To construct the coupling, run the pair of chains with transitions (21.10) until they meet. Afterwards, keep them together. To obtain (21.9), note that if the chain (X_t, Y_t) is started with the distribution $\delta_x \times \pi$, then for fixed t the pair of random variables X_t and Y_t is a coupling of $P^t(x,\cdot)$ with π. Thus by Proposition 4.7 we have

$$\left\|P^t(x,\cdot) - \pi\right\|_{\text{TV}} \le \mathbf{P}_{\delta_x\times\pi}\{X_t \ne Y_t\} \le \mathbf{P}_{\delta_x\times\pi}\{\tau > t\}. \tag{21.12}$$

From (21.11),

$$\lim_{t\to\infty} \mathbf{P}_{\delta_x\times\pi}\{\tau > t\} = \sum_{y\in\Omega} \pi(y) \lim_{t\to\infty} \mathbf{P}_{x,y}\{\tau > t\} = 0.$$

This and (21.12) imply (21.9). ■

EXAMPLE 21.15. Consider a nearest-neighbor random walk on $\mathbb{Z}^+$ which moves up with probability p and down with probability q. If the walk is at 0, it remains at 0 with probability q. Assume that $q > p$.

The equation $\pi = \pi P$ reads as

$$\pi(0) = q\pi(1) + q\pi(0),$$
$$\pi(k) = p\pi(k-1) + q\pi(k+1).$$

Solving, $\pi(1) = \pi(0)(p/q)$ and working up the ladder,

$$\pi(k) = (p/q)^k \pi(0).$$

Here π can be normalized to be a probability distribution, in which case

$$\pi(k) = (p/q)^k(1 - p/q).$$

Since there is a solution to $\pi P = \pi$ which is a probability distribution, the chain is positive recurrent.

By Proposition 1.19, if a solution can be found to the detailed balance equations

$$\pi(x)P(x,y) = \pi(y)P(y,x), \quad x, y \in \Omega,$$

then provided π is a probability distribution, the chain is positive recurrent.

EXAMPLE 21.16 (Birth-and-death chains). A ***birth-and-death*** chain on $\{0, 1, \dots, \}$ is a nearest-neighbor chain which moves up when at k with probability p_k and down with probability $q_k = 1 - p_k$. The detailed balance equations are, for $j \ge 1$,

$$\pi(j)p_j = \pi(j+1)q_{j+1}.$$

Thus $\pi(j+1)/\pi(j) = p_j/q_{j+1}$ and so

$$\pi(k) = \pi(0)\prod_{j=0}^{k-1}\frac{\pi(j+1)}{\pi(j)} = \pi(0)\prod_{j=0}^{k-1}\frac{p_j}{q_{j+1}}.$$

This can be made into a probability distribution provided that

$$\sum_{k=1}^{\infty}\prod_{j=0}^{k-1}\frac{p_j}{q_{j+1}} < \infty, \tag{21.13}$$

in which case we take $\pi(0)^{-1}$ to equal this sum.

If the sum in (21.13) is finite, the chain is positive recurrent.

21.4. Null Recurrence and Convergence

We now discuss the asymptotic behavior of $P^t(x,y)$ in the null recurrent case.

THEOREM 21.17. *If P is the transition matrix on Ω of a null-recurrent irreducible chain, then*

$$\lim_{t\to\infty} P^t(x,y) = 0 \quad \text{for all } x,y\in\Omega. \tag{21.14}$$

PROOF. *Step 1.* It is sufficient to prove $P^t(x,x)\to 0$ for a fixed state x: why? Given $x,y\in\Omega$, by irreducibility, there exists k such that $P^k(y,x)>0$. Since

$$P^{t+k}(x,x) \geq P^t(x,y)P^k(y,x),$$

$P^t(x,x)\to 0$ implies $P^t(x,y)\to 0$.

Step 2. It is sufficient to prove (21.14) for aperiodic P: why? Fix $x\in\Omega$, let $\ell := \gcd\{t : P^t(x,x)>0\}$, and let

$$\tilde{X} := \{y : \text{there exists } k \text{ with } P^{\ell k}(x,y)>0\}.$$

Then P^ℓ is an irreducible aperiodic transition matrix on $\tilde{X}$. Thus we may and shall assume that the original matrix P is irreducible and aperiodic.

Step 3. The measure $\tilde{\pi}$ defined by

$$\tilde{\pi}(y) = \mathbf{E}_x\left(\sum_{t=0}^{\tau_x^+-1}\mathbf{1}_{\{X_t=y\}}\right) \tag{21.15}$$

is a stationary measure. This was shown in the proof of Proposition 1.14 for finite state spaces, and the proof works nearly the same way for countable state spaces. We note that $\tilde{\pi}(y)<\infty$ for all $y\in\Omega$. Why? If the walk visits y before returning to x, the number of additional visits to y before hitting x is a geometric random variable with parameter $\mathbf{P}_y\{\tau_y<\tau_x\}<1$. Note also that $\tilde{\pi}(x)=1$. By null recurrence, $\tilde{\pi}(\Omega)=\infty$.

Step 4. Given M, find a finite set $F\subset\Omega$ with $\tilde{\pi}(F)\geq M$ and consider the conditional distribution μ_F defined by

$$\mu_F(A) = \frac{\tilde{\pi}(A\cap F)}{\tilde{\pi}(F)}.$$

We have

$$\mu_F P^t(x) = \sum_{y\in\Omega}\mu_F(y)P^t(y,x) \leq \frac{1}{\tilde{\pi}(F)}\sum_{y\in\Omega}\tilde{\pi}(y)P^t(y,x) = \frac{\tilde{\pi}P^t(x)}{\tilde{\pi}(F)} = \frac{\tilde{\pi}(x)}{\tilde{\pi}(F)} \leq \frac{1}{M}.$$

By irreducibility and aperiodicity, for all $y \in F$ there exists m_y such that for all $t \geq m_y$ we have $P^t(x,y) > 0$. Let $m := \max_{y \in F} m_y$ and $\varepsilon := \min_{y \in F} P^m(x,y)/\mu_F(y) > 0$.

Step 5. Define a probability measure ν on Ω by

$$P^m(x,\cdot) = \varepsilon \mu_F + (1-\varepsilon)\nu. \tag{21.16}$$

By recurrence,

$$\lim_{N \to \infty} \mathbf{P}_\nu\{\tau_x^+ > N\} = 0.$$

Choose N with $\mathbf{P}_\nu\{\tau_x^+ > N\} \leq \varepsilon/M$. Observe that

$$\mathbf{P}_\nu\{X_t = x\} = \nu P^t(x) \leq \sum_{k=1}^{N} \mathbf{P}_\nu\{\tau_x^+ = k\} P^{t-k}(x,x) + \mathbf{P}_\nu\{\tau_x^+ > N\}.$$

Thus

$$\limsup_{t \to \infty} \mathbf{P}_\nu\{X_t = x\} \leq \limsup_{t \to \infty} P^t(x,x) + \frac{\varepsilon}{M}.$$

Since, by (21.16), for all $t \geq m$

$$P^t(x,x) \leq \varepsilon \mu_F P^{t-m}(x) + (1-\varepsilon)\nu P^{t-m}(x),$$

we conclude that

$$\limsup_{t \to \infty} P^t(x,x) \leq \frac{\varepsilon}{M} + (1-\varepsilon)\left(\limsup_{t \to \infty} P^t(x,x) + \frac{\varepsilon}{M}\right).$$

Therefore,

$$\limsup_{t \to \infty} P^t(x,x) \leq \frac{2}{M}.$$

Since M is arbitrary, the proof is complete. ∎

21.5. Bounds on Return Probabilities

The following is from Barlow, Coulhon, and Kumagai (2005) (cf. Proposition 3.3 there), although the proof given here is different.

THEOREM 21.18. *Let G be an infinite graph with maximum degree at most Δ, and consider the lazy simple random walk on G. For an integer $r > 0$ let $B(x,r)$ denote the ball of radius r (using the graph distance) centered at x. Then for $T = r \cdot |B(x,r)|$ we have*

$$P^T(x,x) \leq \frac{3\Delta^2}{|B(x,r)|}.$$

PROOF. It is clear that in order to prove the statement we may assume we are performing a random walk on the *finite* graph $B(x,T)$ instead of G. Let $(X_t)_{t=0}^\infty$ denote the lazy simple random walk on $B(x,T)$ and denote its stationary distribution by π. Define

$$\tau(x) := \min\{t \geq T : X_t = x\}.$$

We also consider the ***induced chain*** on $B = B(x,r)$ and denote this by $(\widetilde{X}_t)_{t=1}^\infty$. To define it formally, let $\tau_1 < \tau_2 < \cdots$ be all the times such that $X_{\tau_t} \in B$ and write $\widetilde{X}_t = X_{\tau_t}$. We write $\widetilde{\pi}$ for the corresponding stationary distribution on $B = B(x,r)$ and $\widetilde{\tau}(x)$ for the smallest t such that $\tau_t \geq T$ and $\widetilde{X}_t = x$. For any $x \in B$ we have that $\pi(x) = \widetilde{\pi}(x)\pi(B)$. Also, Lemma 10.5 gives that

$$\mathbf{E}_x\,(\text{number of visits of } X_t \text{ to } y \text{ before time } \tau(x)) = \pi(y)\mathbf{E}_x\tau(x).$$

We sum this over $y \in B$ to get

$$\mathbf{E}_x \text{ (number of visits of } X_t \text{ to } B \text{ before time } \tau(x)) = \pi(B)\mathbf{E}_x\tau(x).$$

Observe that the number of visits of X_t to B before $\tau(x)$ equals $\widetilde{\tau}(x)$ and hence

$$\mathbf{E}_x\tau(x) = \frac{\mathbf{E}_x\widetilde{\tau}(x)}{\pi(B)}. \tag{21.17}$$

We now use Lemma 10.5 again to get

$$\begin{aligned}\sum_{t=0}^{T-1} P^t(x,x) &= \mathbf{E}_x \text{ (number of visits to } x \text{ before time } \tau(x)) \\ &= \pi(x)\mathbf{E}_x\tau(x) = \widetilde{\pi}(x)\mathbf{E}_x\widetilde{\tau}(x),\end{aligned} \tag{21.18}$$

where the last equality is due to (21.17). Denote by σ the minimal $t \geq T$ such that $X_t \in B$ and let ν be the distribution of X_σ. Observe that $\mathbf{E}_x\widetilde{\tau}(x) \leq T + \mathbf{E}_\nu\widetilde{\tau}_0(x)$ where $\widetilde{\tau}_0(x)$ is the first hitting time of x in the induced chain. Since $P^t(x,x)$ is weakly decreasing in t (Proposition 10.18), we infer that

$$TP^T(x,x) \leq \widetilde{\pi}(x)[T + \mathbf{E}_\nu\widetilde{\tau}_0(x)].$$

We use the Commute Time Identity (Proposition 10.6) and bound the effective resistance from above by the distance to get

$$\mathbf{E}_\nu\widetilde{\tau}_0(x) \leq 2\Delta r|(B(x,r))|.$$

Since $\widetilde{\pi}(x) \leq \Delta/|B(x,r)|$, we conclude that

$$TP^T(x,x) \leq \frac{\Delta T}{|B(x,r)|} + 2\Delta^2 r.$$

This immediately gives that

$$P^T(x,x) \leq \frac{\Delta}{|B(x,r)|} + \frac{2\Delta^2 r}{T}.$$

Recalling that $T = r|B(x,r)|$ finishes the proof. ■

Exercises

Exercise 21.1. Use the Strong Law of Large Numbers to give a proof that the biased random walk in Example 21.2 is transient.

Exercise 21.2. Suppose that P is irreducible. Show that if $\pi = \pi P$ for a probability distribution π, then $\pi(x) > 0$ for all $x \in \Omega$.

Exercise 21.3. Fix $k > 1$. Define the ***k-fuzz*** of an undirected graph $G = (V, E)$ as the graph $G_k = (V, E_k)$ where for any two distinct vertices $v, w \in V$, the edge $\{v, w\}$ is in E_k if and only if there is a path of at most k edges in E connecting v to w. Show that for G with bounded degrees, G is transient if and only if G_k is transient.

A solution can be found in Doyle and Snell (1984, Section 8.4).

Exercise 21.4. Show that any subgraph of a recurrent graph must be recurrent.

EXERCISE 21.5. Let P be an irreducible and aperiodic transition matrix on Ω. Let $\tilde{P}$ be the matrix on $\Omega \times \Omega$ defined by

$$\tilde{P}((x,y),(z,w)) = P(x,z)P(y,z), \quad (x,y) \in \Omega \times \Omega,\ (z,w) \in \Omega \times \Omega.$$

Show that $\tilde{P}$ is irreducible.

EXERCISE 21.6. Consider the discrete-time single server FIFO (first in, first out) queue: at every step, if there is a customer waiting, exactly one of the following happens:

(1) a new customer arrives (with probability α) or
(2) an existing customer is served (with probability $\beta = 1 - \alpha$).

If there are no customers waiting, then (1) still has probability α, but (2) is replaced by "nothing happens". Let X_t be the number of customers in the queue at time t.

Show that (X_t) is

(a) positive recurrent if $\alpha < \beta$,
(b) null recurrent if $\alpha = \beta$,
(c) transient if $\alpha > \beta$.

EXERCISE 21.7. Consider the same set-up as Exercise 21.6. In the positive recurrent case, determine the stationary distribution π and the π-expectation of the time T from the arrival of a customer until he is served.

REMARK 21.19. In communication theory one talks of *packets* instead of customers.

EXERCISE 21.8. Let P be the transition matrix for simple random walk on $\mathbb{Z}$. Show that the walk is not positive recurrent by showing there are no probability distributions π on $\mathbb{Z}$ satisfying $\pi P = \pi$.

Notes

Further reading. Many texts, including Feller (1968) and Doyle and Snell (1984), also give proofs of the recurrence of random walk in one and two dimensions and of the transience in three or more.

Lyons (1983) used flows for analyzing chains with infinite state spaces.

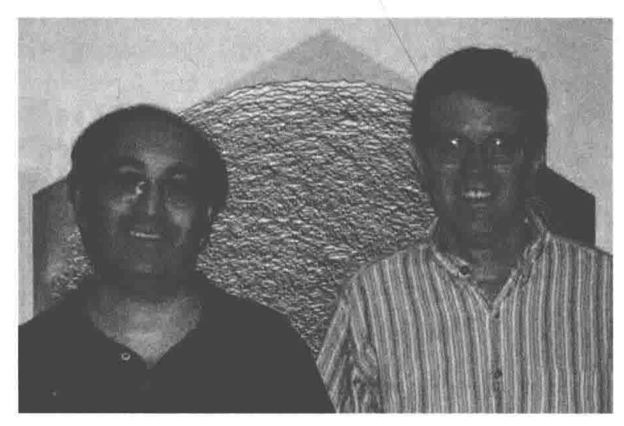
For much more on infinite networks, see Soardi (1994), Woess (2000), and Lyons and Peres (2008).

For more on Markov chains with infinite state spaces, see, e.g., Feller (1968), Norris (1998), or Kemeny, Snell, and Knapp (1976). See also Thorisson (2000).

CHAPTER 22

Coupling from the Past

by James G. Propp and David B. Wilson



J.G. Propp (left) and D.B. Wilson (right).

22.1. Introduction

In Markov chain Monte Carlo studies, one attempts to sample from a probability distribution π by running a Markov chain whose unique stationary distribution is π. Ideally, one has proved a theorem that guarantees that the time for which one plans to run the chain is substantially greater than the mixing time of the chain, so that the distribution $\tilde{\pi}$ that one's procedure actually samples from is known to be close to the desired π in variation distance. More often, one merely hopes that this is the case, and the possibility that one's samples are contaminated with substantial initialization bias cannot be ruled out with complete confidence.

The "coupling from the past" (CFTP) procedure introduced by Propp and Wilson (1996) provides one way of getting around this problem. Where it is applicable, this method determines on its own how long to run and delivers samples that are governed by π itself, rather than $\tilde{\pi}$. Many researchers have found ways to apply the basic idea in a wide variety of settings (see `http://dbwilson.com/exact/` for pointers to this research). Our aim here is to explain the basic method and to give a few of its applications.

It is worth stressing at the outset that CFTP is especially valuable as an alternative to standard Markov chain Monte Carlo when one is working with Markov chains for which one suspects, but has not proved, that rapid mixing occurs. In such cases, the availability of CFTP makes it less urgent that theoreticians obtain bounds on the mixing time, since CFTP (unlike Markov chain Monte Carlo) cleanly separates the issue of efficiency from the issue of quality of output. That is

to say, one's samples are guaranteed to be uncontaminated by initialization bias, regardless of how quickly or slowly they are generated.

Before proceeding, we mention that there are other algorithms that may be used for generating perfect samples from the stationary distribution of a Markov chain, including Fill's algorithm (Fill, 1998; Fill, Machida, Murdoch, and Rosenthal, 2000), "dominated CFTP" (Kendall and Møller, 2000), "read-once CFTP" (Wilson, 2000b), and the "randomness recycler" (Fill and Huber, 2000). Each of these has its merits, but since CFTP is conceptually the simplest of these, it is the one that we shall focus our attention on here.

As a historical aside, we mention that the conceptual ingredients of CFTP were in the air even before the versatility of the method was made clear in Propp and Wilson (1996). Precursors include Letac (1986), Thorisson (1988), and Borovkov and Foss (1992). Even back in the 1970's, one can find foreshadowings in the work of Ted Harris (on the contact process, the exclusion model, random stirrings, and coalescing and annihilating random walks), David Griffeath (on additive and cancellative interacting particle systems), and Richard Arratia (on coalescing Brownian motion). One can even see traces of the idea in the work of Loynes (1962) forty-five years ago. See also the survey by Diaconis and Freedman (1999).

22.2. Monotone CFTP

The basic idea of coupling from the past is quite simple. Suppose that there is an ergodic Markov chain that has been running either forever or for a very long time, long enough for the Markov chain to have reached (or very nearly reached) its stationary distribution. Then the state that the Markov chain is currently in is a sample from the stationary distribution. If we can figure out what that state is, by looking at the recent randomizing operations of the Markov chain, then we have a sample from its stationary distribution. To illustrate these ideas, we show how to apply them to the Ising model of magnetism (introduced in Section 3.3.5 and studied further in Chapter 15).

Recall that an Ising system consists of a collection of n interacting spins, possibly in the presence of an external field. Each spin may be aligned up or down. Spins that are close to each other prefer to be aligned in the same direction, and all spins prefer to be aligned with the external magnetic field (which sometimes varies from site to site). These preferences are quantified in the total energy H of the system

$$H(\sigma) = -\sum_{i<j} \alpha_{i,j}\sigma_i\sigma_j - \sum_i B_i\sigma_i,$$

where B_i is the strength of the external field as measured at site i, σ_i is 1 if spin i is aligned up and -1 if it is aligned down, and $\alpha_{i,j} \geq 0$ represents the interaction strength between spins i and j. The probability of a given spin configuration is given by $Z^{-1}\exp[-\beta H(\sigma)]$ where β is the "inverse temperature" and Z is the "partition function," i.e., the normalizing constant that makes the probabilities add up to 1. Often the n spins are arranged in a two-dimensional or three-dimensional lattice, and $\alpha_{i,j} = 1$ if spins i and j are adjacent in the lattice, and $\alpha_{i,j} = 0$ otherwise. The Ising model has been used to model certain substances such as crystals of $FeCl_2$ and $FeCO_3$ and certain phases of carbon dioxide, xenon, and brass — see Baxter (1982) for further background.

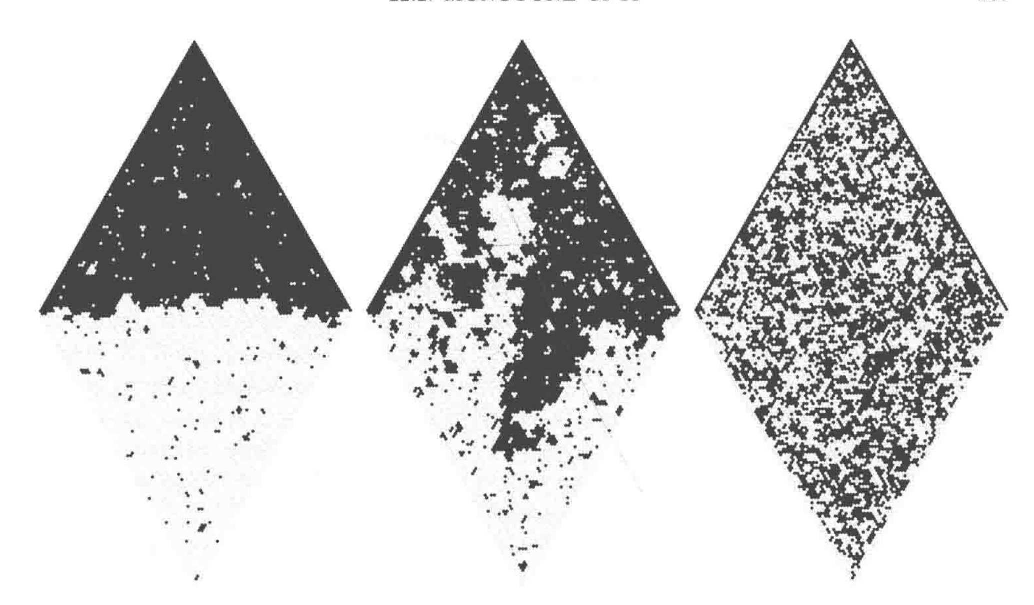


FIGURE 22.1. The Ising model at three different temperatures (below, at, and above the "critical" temperature). Here the spins lie at the vertices of the triangular lattice and are shown as black or white hexagons. The spins along the upper boundaries were forced to be black and the spins along lower boundaries were forced to be white (using an infinite magnetic field on these boundary spins).

We may use the single-site heat bath algorithm, also known as Glauber dynamics, to sample Ising spin configurations. (Glauber dynamics was introduced in Section 3.3.) A single move of the heat-bath algorithm may be summarized by a pair of numbers (i, u), where i represents a spin site (say that i is a uniformly random site), and u is a uniformly random real number between 0 and 1. The heat-bath algorithm randomizes the alignment of spin i, holding all of the remaining magnets fixed, and uses the number u when deciding whether the new spin should be up or down. There are two possible choices for the next state, denoted by $\sigma_\uparrow$ and $\sigma_\downarrow$. We have $\Pr[\sigma_\uparrow]/\Pr[\sigma_\downarrow] = e^{-\beta(H(\sigma_\uparrow)-H(\sigma_\downarrow))} = e^{-\beta(\Delta H)}$. The update rule is that the new spin at site i is up if $u < \Pr[\sigma_\uparrow]/(\Pr[\sigma_\uparrow]+\Pr[\sigma_\downarrow])$, and otherwise the new spin is down. It is easy to check that this defines an ergodic Markov chain with the desired stationary distribution.

Recall our supposition that the randomizing process, in this case the single-site heat bath, has been running for all time. Suppose that someone has recorded all the randomizing operations of the heat bath up until the present time. They have not recorded what the actual spin configurations or Markov chain transitions are, but merely which sites were updated and which random number was used to update the spin at the given site. Given this recorded information, our goal is to determine the state of the Markov chain at the present time (time 0), since, as we have already determined, this state is a sample from the stationary distribution of the Markov chain.

To determine the state at time 0, we make use of a natural partial order with which the Ising model is equipped: we say that two spin-configurations σ and τ satisfy $\sigma \preceq \tau$ when each spin-up site in σ is also spin-up in τ. Notice that if $\sigma \preceq \tau$

and we update both σ and τ with the same heat-bath update operation (i, u), then because site i has at least as many spin-up neighbors in τ as it does in σ and because of our assumption that the $\alpha_{i,j}$'s are nonnegative, we have $\Pr[\tau_\uparrow]/\Pr[\tau_\downarrow] \geq \Pr[\sigma_\uparrow]/\Pr[\sigma_\downarrow]$, and so the updated states σ' and τ' also satisfy $\sigma' \preceq \tau'$. (We say that the randomizing operation respects the partial order $\preceq$.) Notice also that the partial order $\preceq$ has a maximum state $\hat{1}$, which is spin-up at every site, and a minimum state $\hat{0}$, which is spin-down at every site.

This partial order enables us to obtain upper and lower bounds on the state at the present time. We can look at the last T randomizing operations, figure out what would happen if the Markov chain were in state $\hat{1}$ at time $-T$, and determine where it would be at time 0. Since the Markov chain is guaranteed to be in a state which is $\preceq \hat{1}$ at time $-T$ and since the randomizing operations respect the partial order, we obtain an upper bound on the state at time 0. Similarly we can obtain a lower bound on the state at time 0 by applying the last T randomizing operations to the state $\hat{0}$. It could be that we are lucky and the upper and lower bounds are equal, in which case we have determined the state at time 0. If we are not so lucky, we could look further back in time, say at the last $2T$ randomizing operations, and obtain better upper and lower bounds on the state at the present time. So long as the upper and lower bounds do not coincide, we can keep looking further and further back in time (see Figure 22.2). Because the Markov chain is ergodic, when it is started in $\hat{1}$ and T is large enough, there is some positive chance that it will reach $\hat{0}$, after which the upper and lower bounds are guaranteed to coincide. In the limit as $T \to \infty$, the probability that the upper and lower bounds agree at time 0 tends to 1, so almost surely we eventually succeed in determining the state at time 0.

The randomizing operation (the heat-bath in the above Ising model example) defines a (grand) coupling of the Markov chain, also sometimes called a *stochastic flow,* since it couples Markov chains started from all the states in the state space. (Grand couplings were discussed in Section 5.4.) For CFTP, the choice of the coupling is as important as the choice of the Markov chain. To illustrate this, we consider another example, tilings of a regular hexagon by lozenges, which are

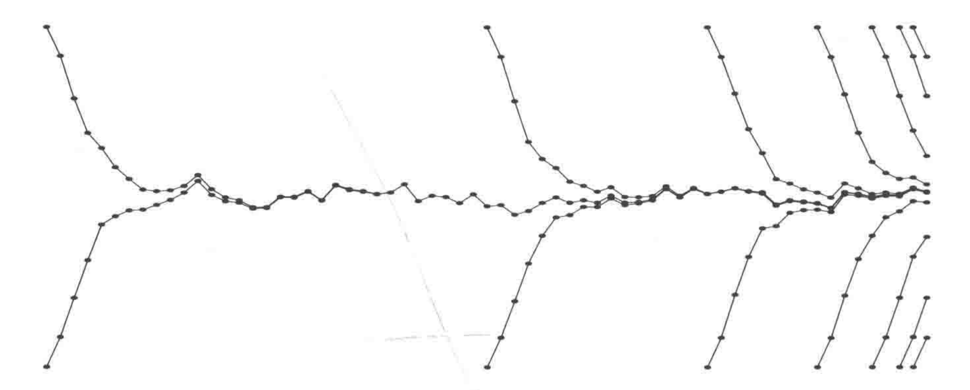


FIGURE 22.2. Illustration of CFTP in the monotone setting. Shown are the heights of the upper and lower trajectories started at various starting times in the past. When a given epoch is revisited later by the algorithm, it uses the same randomizing operation.

$60°/120°$ rhombuses (see Figure 22.3). The set of lozenge tilings comes equipped

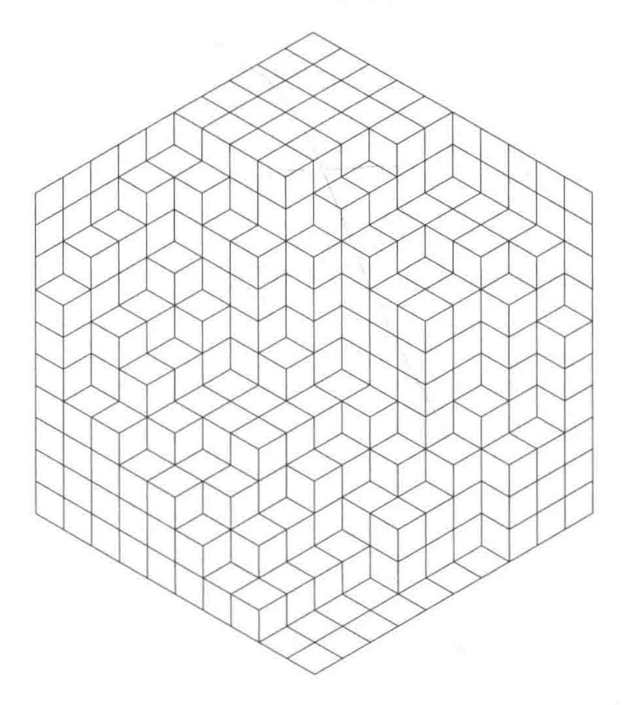


FIGURE 22.3. Tilings of a regular hexagon by lozenges. Alternatively, these tilings may be viewed three-dimensionally, as a collection of little three-dimensional boxes sitting within a larger box.

with a natural partial order $\preceq$: we say that one tiling lies below another tiling if, when we view the tilings as collections of little three-dimensional boxes contained within a large box, the first collection of boxes is a subset of the other collection of boxes. The minimum configuration $\hat{0}$ is just the empty collection of little boxes, and the maximum configuration $\hat{1}$ is the full collection of little boxes.

A site in the tiling is just a vertex of one of the rhombuses that is contained within the interior of the hexagon. For each possible tiling, these sites form a triangular lattice. If a site is surrounded by exactly three lozenges, then the three lozenges will have three different orientations, one of which is horizontal if the regular hexagon is oriented as shown in Figure 22.3. There are two different ways for a site to be surrounded by three lozenges — the horizontal lozenge will lie either above the site or below it. One possible randomizing operation would with probability $1/2$ do nothing and with probability $1/2$ pick a uniformly random site in the tiling, and if that site is surrounded by three lozenges, rearrange those three lozenges. Another possible randomizing operation would pick a site uniformly at random and then if the site is surrounded by three lozenges, with probability $1/2$ arrange the three lozenges so that the horizontal one is below the site and with probability $1/2$ arrange them so that the horizontal lozenge is above the site. When the tiling is viewed as a collection of boxes, this second randomizing operation either tries to remove or add (with probability $1/2$ each) a little box whose projection into the plane of the tiling is at the site. These attempts to add or remove a little box only succeed when the resulting configuration of little boxes would be stable under

gravity; otherwise the randomizing operation leaves the configuration alone. It is straightforward to check that both of these randomizing operations give rise to the same Markov chain, i.e., a given tiling can be updated according to the first randomizing operation or the second randomizing operation, and either way, the distribution of the resulting tiling will be precisely the same. However, for purposes of CFTP the second randomizing operation is much better, because it respects the partial order $\preceq$, whereas the first randomizing operation does not.

With the Ising model and tiling examples in mind, we give pseudocode for "monotone CFTP," which is CFTP when applied to state spaces with a partial order $\preceq$ (with a top state $\hat{1}$ and bottom state $\hat{0}$) that is preserved by the randomizing operation:

```
T ← 1
repeat
    upper ← 1̂
    lower ← 0̂
    for t = −T to −1
        upper ← φ(upper, U_t)
        lower ← φ(lower, U_t)
    T ← 2T
until upper = lower
return upper
```

Here the variables U_t represent the intrinsic randomness used in the randomizing operations. In the Ising model heat-bath example above, U_t consists of a random number representing a site together with a random real number between 0 and 1. In the tiling example, U_t consists of the random site together with the outcome of a coin toss. The procedure φ deterministically updates a state according to the random variable U_t.

Recall that we are imagining that the randomizing operation has been going on for all time, that someone has recorded the random variables U_t that drive the randomizing operations, and that our goal is to determine the state at time 0. Clearly if we read the random variable U_t more than one time, it would have the same value both times. Therefore, when the random mapping $\varphi(\cdot, U_t)$ is used in one iteration of the repeat loop, for any particular value of t, it is essential that the same mapping be used in all subsequent iterations of the loop. We may accomplish this by storing the U_t's; alternatively, if (as is typically the case) our U_t's are given by some pseudo-random number generator, we may simply suitably reset the random number generator to some specified seed $seed(i)$ each time t equals -2^i.

REMARK 22.1. Many people ask about different variations of the above procedure, such as what happens if we couple into the future or what happens if we use fresh randomness each time we need to refer to the random variable U_t. There is a simple example that rules out the correctness of all such variations that have been suggested. Consider the state space $\{1, 2, 3\}$, where the randomizing operation with probability $1/2$ increments the current state by 1 (unless the state is 3) and with probability $1/2$ decrements the current state by 1 (unless the state is 1). We leave it as an exercise to verify that this example rules out the correctness of the above two variants. There are in fact other ways to obtain samples from the stationary distribution of a monotone Markov chain, such as by using Fill's algorithm (Fill, 1998) or "read-once CFTP" (Wilson, 2000b), but these are not the sort of procedures that one will discover by randomly mutating the above procedure.

It is worth noting that monotone CFTP is efficient whenever the underlying Markov chain is rapidly mixing. Propp and Wilson (1996) proved that the number of randomizing operations that monotone CFTP performs before returning a sample is at least t_{mix} and at most $O(t_{\text{mix}} \log H)$, where t_{mix} is the mixing time of the Markov chain when measured with the total variation distance and H denotes the length of the longest totally ordered chain of states between $\hat{0}$ and $\hat{1}$.

There are a surprisingly large number of Markov chains for which monotone CFTP may be used (see Propp and Wilson (1996) and other articles listed in `http://dbwilson.com/exact/`). In the remainder of this chapter we describe a variety of scenarios in which CFTP has been used even when monotone CFTP cannot be used.

22.3. Perfect Sampling via Coupling from the Past

Computationally, one needs three things in order to be able to implement the CFTP strategy: a way of generating (and representing) certain maps from the state space Ω to itself; a way of composing these maps; and a way of ascertaining whether *total coalescence* has occurred, i.e., a way of ascertaining whether a certain composite map (obtained by composing many random maps) collapses all of Ω to a single element.

The first component is what we call the random map procedure; we model it as an oracle that on successive calls returns independent, identically distributed functions f from Ω to Ω, governed by some selected probability distribution P (typically supported on a very small subset of the set of all maps from Ω to itself). We use the oracle to choose independent, identically distributed maps f_{-1}, f_{-2}, f_{-3}, $\ldots$, f_{-T}, where how far into the past we have to go (T steps) is determined during run-time itself. (In the notation of the previous section, $f_t(x) = \varphi(x, U_t)$. These random maps are also known as grand couplings, which were discussed in Section 5.4.) The defining property that T must have is that the composite map

$$F_{-T}^{0} \stackrel{\text{def}}{=} f_{-1} \circ f_{-2} \circ f_{-3} \circ \cdots \circ f_{-T}$$

must be collapsing. Finding such a T thus requires that we have both a way of composing f's and a way of testing when such a composition is collapsing. (Having the test enables one to find such a T, since one can iteratively test ever-larger values of T, say by successive doubling, until one finds a T that works. Such a T will be a random variable that is measurable with respect to $f_{-T}, f_{-T+1}, \ldots, f_{-1}$.)

Once a suitable T has been found, the algorithm outputs $F_{-T}^{0}(x)$ for any $x \in \Omega$ (the result will not depend on x, since F_{-T}^{0} is collapsing). We call this output the CFTP sample. It must be stressed that when one is attempting to determine a usable T by guessing successively larger values and testing them in turn, one must use the *same* respective maps f_i during each test. That is, if we have just tried starting the chain from time $-T_1$ and failed to achieve coalescence, then, as we proceed to try starting the chain from time $-T_2 < -T_1$, we must use the same maps f_{-T_1}, f_{-T_1+1}, $\ldots$, f_{-1} as in the preceding attempt. This procedure is summarized below:

$T \leftarrow 1$
while $f_{-1} \circ \cdots \circ f_{-T}$ is not totally coalescent
    increase T
return the value to which $f_{-1} \circ \cdots \circ f_{-T}$ collapses Ω

Note that the details of how one increases T affect the computational efficiency of the procedure but not the distribution of the output; in most applications it is most natural to double T when increasing it (as in Sections 22.2 and 22.4), but sometimes it is more natural to increment T when increasing it (as in Section 22.5).

As long as the nature of P guarantees (almost sure) eventual coalescence, and as long as P bears a suitable relationship to the distribution π, the CFTP sample will be distributed according to π. Specifically, it is required that P preserve π in the sense that if a random state x is chosen in accordance with π and a random map f is chosen in accordance with P, then the state $f(x)$ will be distributed in accordance with π. In the next several sections we give examples.

22.4. The Hardcore Model

Recall from Section 3.3.4 that the states of the hardcore model are given by subsets of the vertex set of a finite graph G, or equivalently, by 0, 1-valued functions on the vertex set. We think of 1 and 0 as respectively denoting the presence or absence of a particle. In a legal state, no two adjacent vertices may both be occupied by particles. The probability of a particular legal state is proportional to λ^m, where m is the number of particles (which depends on the choice of state) and λ is some fixed parameter value. We denote this probability distribution by π. That is, $\pi(\sigma) = \lambda^{|\sigma|}/Z$ where σ is a state, $|\sigma|$ is the number of particles in that state, and $Z = \sum_\sigma \lambda^{|\sigma|}$. Figure 22.4 shows some hardcore states for different values of λ when the graph G is the toroidal grid.

The natural single-site heat-bath Markov chain for hardcore states would pick a site at random, forget whether or not there is a particle at that site, and then place a particle at the site with probability $\lambda/(\lambda+1)$ if there are no neighboring particles or with probability 0 if there is a neighboring particle.

For general (non-bipartite) graphs G there is no monotone structure which would allow one to use monotone CFTP. But Häggström and Nelander (1999) and Huber (1998) proposed the following scheme for using CFTP with the single-site heat-bath Markov chain. One can associate with each *set* of hardcore states a three-valued function on the vertex set, where the value "1" means that all states in the set are known to have a particle at that vertex, the value "0" means that all states in the set are known to have a vacancy at that vertex, and the value "?" means that it is possible that some of the states in the set have a particle there while others have a vacancy. Initially we place a "?" at every site since the Markov chain could be in any state. We can operate directly on this three-valued state-model by means of simple rules that mimic the single-site heat-bath. The randomizing operation picks a random site and proposes to place a particle there with probability $\lambda/(\lambda+1)$ or proposes to place a vacancy there with probability $1/(\lambda+1)$. Any proposal to place a vacancy always succeeds for any state in the current set, so in this case a "0" is placed at the site. A proposal to place a particle at the site succeeds only if no neighboring site has a particle, so in this case we place a "1" if all neighboring sites have a "0", and otherwise we place a "?" at the site since the proposal to place a particle there may succeed for some states in the set and fail for other states. After the update, the "0, 1, ?" configuration describes any possible state that the Markov chain may be in after the single-site heat-bath operation. It is immediate that if the "0, 1, ?" Markov chain, starting from the all-?'s state, ever reaches a state in which there are no ?'s, then the single-site heat-bath chain, using the same random

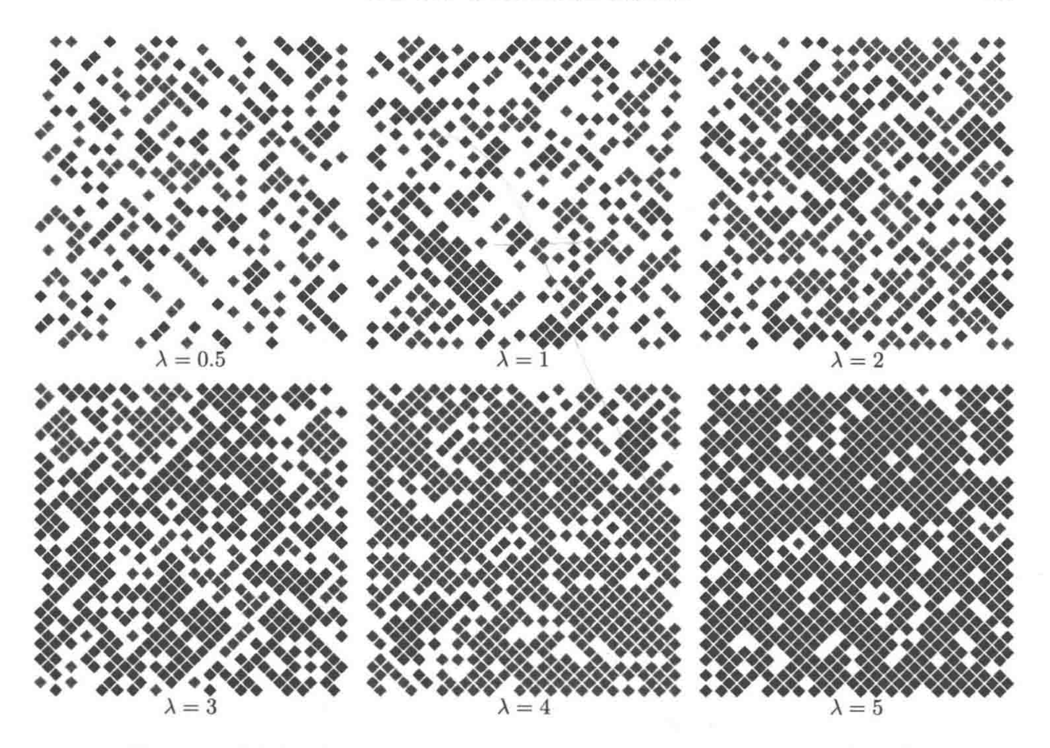


FIGURE 22.4. Hardcore model on the 40×40 square grid with periodic boundary conditions, for different values of λ. Particles are shown as diamonds, and the constraint that no two particles are adjacent is equivalent to the constraint that no two diamonds overlap. Particles on the even sublattice (where the x-coordinate and y-coordinate have the same parity) are shown in dark gray, and particles on the odd sublattice are shown in light gray. There is a critical value of λ above which the hardcore model typically has a majority of particles on one of these two sublattices. CFTP generates random samples for values of λ beyond those for which Glauber dynamics is currently known to be rapidly mixing.

proposals, maps all initial states into the same final state. Hence we might want to call the "0, 1, ?" Markov chain the "certification chain," for it tells us when the stochastic flow of primary interest has achieved coalescence.

One might fear that it would take a long time for the certification chain to certify coalescence, but Häggström and Nelander (1999) show that the number of ?'s tends to shrink to zero exponentially fast provided $\lambda < 1/\Delta$, where Δ is the maximum degree of the graph. Recall from Theorem 5.8 that the Glauber dynamics Markov chain is rapidly mixing when $\lambda < 1/(\Delta - 1)$ — having the number of ?'s shrink to zero rapidly is a stronger condition than rapid mixing. The best current bounds for general graphs is that Glauber dynamics is rapidly mixing if $\lambda \leq 2/(\Delta - 2)$ (Vigoda, 2001; Dyer and Greenhill, 2000). For particular graphs of interest, such as the square lattice, in practice the number of ?'s shrinks to zero rapidly for values of λ much larger than what these bounds guarantee. Such observations constitute empirical evidence in favor of rapid mixing for larger λ's.

22.5. Random State of an Unknown Markov Chain

Now we come to a problem that in a sense encompasses all the cases we have discussed so far: the problem of sampling from the stationary distribution π of a general Markov chain. Of course, in the absence of further strictures this problem admits a trivial "solution": just solve for the stationary distribution analytically! In the case of the systems studied in Sections 22.2 and 22.4, this is not practical, since the state spaces are large. We now consider what happens if the state space is small but the analytic method of simulation is barred by imposing the constraint that the transition probabilities of the Markov chain are unknown: one merely has access to a black box that simulates the transitions.

It might seem that, under this stipulation, no solution to the problem is possible, but in fact a solution was found by Asmussen, Glynn, and Thorisson (1992). However, their algorithm was not very efficient. Subsequently Aldous (1995) and Lovász and Winkler (1995a) found faster procedures (although the algorithm of Aldous involves controlled but non-zero error). The CFTP-based solution given below is even faster than that of Lovász and Winkler.

For pictorial concreteness, we envision the Markov chain as a biased random walk on some directed graph G whose arcs are labeled with weights, where the transition probabilities from a given vertex are proportional to the weights of the associated arcs (as in the preceding section). We denote the vertex set of G by Ω, and denote the stationary distribution on Ω by π. Propp and Wilson (1998) give a CFTP-based algorithm that lets one sample from this distribution π.

Our goal is to define suitable random maps from Ω to Ω in which many states are mapped into a single state. We might therefore define a random map from Ω to itself by starting at some fixed vertex r, walking randomly for some large number T of steps, and mapping all states in Ω to the particular state v that one has arrived at after T steps. However, v is subject to initialization bias, so this random map procedure typically does not preserve π in the sense defined in Section 22.3.

What actually works is a multi-phase scheme of the following sort: start at some vertex r and take a random walk for a *random* amount of time T_1, ending at some state v; then map every state that has been visited during that walk to v. In the second phase, continue walking from v for a further random amount of time T_2, ending at some new state v'; then map every state that was visited during the second phase but not the first to v'. In the third phase, walk from v' for a random time to a new state v'', and map every hitherto-unvisited state that was visited during that phase to the state v'', and so on. Eventually, every state gets visited, and every state gets mapped to some state. Such maps are easy to compose, and it is easy to recognize when such a composition is coalescent (it maps every state to one particular state).

There are two constraints that our random durations T_1, T_2, ... must satisfy if we are planning to use this scheme for CFTP. (For convenience we will assume henceforth that the T_i's are i.i.d.) First, the distribution of each T_i should have the property that, at any point during the walk, the (conditional) expected time until the walk terminates does not depend on where one is or how one got there. This ensures that the stochastic flow determined by these random maps preserves π. Second, the time for the walk should be neither so short that only a few states get visited by the time the walk ends nor so long that generating even a single random map takes more time than an experimenter is willing to wait. Ideally, the expected

duration of the walk should be on the order of the cover time for the random walk. Propp and Wilson (1998) show that by using the random walk itself to estimate its own cover time, one gets an algorithm that generates a random state distributed according to π in expected time ≤ 15 times the cover time.

At the beginning of this section, we said that one has access to a black box that simulates the transitions. This is, strictly speaking, ambiguous: does the black box have an "input port" so that we can ask it for a random transition from a specified state? Or are we merely passively observing a Markov chain in which we have no power to intervene? This ambiguity gives rise to two different versions of the problem, of separate interest. Our CFTP algorithm works for both of them.

For the "passive" version of the problem, it is not hard to show that no scheme can work in expected time less than the expected cover time of the walk, so in this setting our algorithm runs in time that is within a constant factor of optimal. It is possible to do better in the active setting, but no good lower bounds are currently known for this case.

Exercise

EXERCISE 22.1. Show that in the special case where the graph is bipartite, there is a natural partial order on the space of hardcore configurations that is preserved by Glauber dynamics and that in this case monotone CFTP and CFTP with the "0, 1, ?" Markov chain are equivalent.

Notes

This chapter is based in part on the expository article "Coupling from the Past: a User's Guide," which appeared in *Microsurveys in Discrete Probability*, volume 41 of the *DIMACS Series in Discrete Mathematics and Computer Science*, published by the AMS, and contains excerpts from the article "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics," which appeared in *Random Structures and Algorithms*, volume 9(1&2):223–252, 1996.

For more on perfectly sampling the spanning trees of a graph, see Anantharam and Tsoucas (1989), Broder (1989), and Aldous (1990). For more examples of perfect sampling, see Häggström and Nelander (1998), Wilson (2000a), and the webpage Wilson (2004b).

CHAPTER 23

Open Problems

This list of questions is not meant to be either novel or comprehensive. The selection of topics clearly reflects the interests of the authors. Aldous and Fill (1999) features open problems throughout the book; several have already been solved. We hope this list will be similarly inspirational.

23.1. The Ising Model

For all of these, assume Glauber dynamics unless another transition mechanism is specified.

QUESTION 1 (Positive boundary conditions). Consider the Ising model on the $n \times n$ grid with the boundary forced to have all positive spins. Show that at any temperature the mixing time is at most polynomial in n. An upper bound on the relaxation time of $e^{n^{1/2+\varepsilon}}$ was obtained by Martinelli (1994). The best upper bounds for $d \geq 3$ were obtained by Sugimine (2002).

QUESTION 2 (Monotonicity). Is the spectral gap of the Ising model on a graph G monotone increasing in temperature? Is the spectral gap of the Ising model monotone decreasing in the addition of edges?

There is a common generalization of these two questions to the ferromagnetic Ising model with inhomogeneous interaction strengths. If for simplicity we absorb the temperature into the interaction strengths, the Gibbs distribution for this model can be defined by

$$\mu(\sigma) = \frac{1}{Z} \exp\left(\sum_{\{u,v\} \in E(G)} J_{u,v}\sigma(u)\sigma(v) \right),$$

where $J_{u,v} > 0$ for all edges $\{u, v\}$. In this model, is it true that on any graph the spectral gap is monotone decreasing in each interaction strength $J_{u,v}$? Nacu (2003) proved this stronger conjecture for the cycle.

Even more generally, we may ask whether for a fixed graph and fixed t the distance $\bar{d}(t)$ is monotone increasing in the individual interaction strengths $J_{u,v}$. (Corollary 12.6 and Lemma 4.11 ensure that this is, in fact, a generalization.)

QUESTION 3 (Lower bounds). Is it true that on an n-vertex graph, the mixing time for the Glauber dynamics for Ising is at least $cn \log n$? This is known for bounded degree families (the constant depends on the maximum degree); see Hayes and Sinclair (2007). We conjecture that on any graph, at any temperature, there is a lower bound of $(1/2 + o(1))n \log n$ on the mixing time.

QUESTION 4 (Block dynamics vs. single site dynamics). Consider block dynamics on a family of finite graphs. If the block sizes are bounded, are mixing times always comparable for block dynamics and single site dynamics? This is true for the relaxation times, via comparison of Dirichlet forms.

QUESTION 5 (Systematic updates vs. random updates). Fix a permutation α of the vertices of an n-vertex graph and successively perform Glauber updates at $\alpha(1), \ldots, \alpha(n)$. Call the transition matrix of the resulting operation P_α. That is, P_α corresponds to doing a full sweep of all the vertices. Let P be the transition matrix of ordinary Glauber dynamics.

(i) Does there exist a constant C such that

$$n t_{\mathrm{mix}}(P_\alpha) \leq C t_{\mathrm{mix}}(P)?$$

(ii) Does there exist a constant c such that

$$n t_{\mathrm{mix}}(P_\alpha) \geq c \frac{t_{\mathrm{mix}}(P)}{\log n}?$$

Although theorems are generally proved about random updates, in practice systematic updates are often used for running simulations. (Note that at infinite temperature, a single systematic sweep suffices.) See Dyer, Goldberg, and Jerrum (2006a) and (2006b) for analysis of systematic swap algorithms for colorings.

QUESTION 6 (Ising on transitive graphs). For the Ising model on transitive graphs, is the relaxation time of order n if and only if the mixing time is of order $n \log n$ (as the temperature varies)? This is known to be true for the two-dimensional torus. See Martinelli (1999) for more on what is known on lattices.

23.2. Cutoff

QUESTION 7 (Transitive graphs of bounded degree). Given a sequence of transitive graphs of degree $\Delta \geq 3$, must the family of lazy random walks on these graphs have a cutoff?

QUESTION 8 (Cutoff for Ising on transitive graphs). Consider the Ising model on a transitive graph, e.g. a d-dimensional torus, at high temperature. Is there a cutoff whenever the mixing time is of order $n \log n$? Is this true, in particular, for the cycle? Levin, Luczak, and Peres (2007) showed that the answer is "yes" for the complete graph.

QUESTION 9 (Card shuffling). Do the following shuffling chains have cutoff? All are known to have pre-cutoff.

(a) Random adjacent transpositions (pre-cutoff follows from (16.4) and (16.7)).
(b) Cyclic-to-random transpositions (see Mossel, Peres, and Sinclair (2004)).
(c) Random-to-random insertions (see the thesis of Uyemura-Reyes (2002)). In this shuffle, a card is chosen uniformly at random, removed from the deck, and reinserted into a uniform random position. The other cards retain their original relative order.

QUESTION 10 (Lamplighter on tori). Does the lamplighter on tori of dimension $d \geq 3$ have a cutoff? If there is a total variation cutoff, at what multiple of the cover time of the torus does it occur?

QUESTION 11. Let $(X_t^{(n)})$ denote a family of irreducible reversible Markov chains, either in continuous-time or in lazy discrete-time. Is it true that there is cutoff *in separation distance* if and only if there is cutoff *in total variation distance*? That this is true for birth-and-death chains follows from combining results in Ding, Lubetzky, and Peres (2008b) and Diaconis and Saloff-Coste (2006). A positive answer to this question for lamplighter walks would also answer Question 10, in view of Theorem 19.7.

23.3. Other Problems

QUESTION 12 (Spectral gap of the interchange process). Place a pebble at each vertex of a graph G, and on each edge place an alarm clock that rings at each point of a Poisson process with density 1. When the clock on edge $\{u, v\}$ rings, interchange the pebbles at u and v. This process is called the *interchange process* on G. Handjani and Jungreis (1996) showed that for trees, the interchange process on G and the continuous-time simple random walk on G have the same spectral gap. Is this true for all graphs? This question was raised by Aldous and Diaconis (see Handjani and Jungreis (1996)).

QUESTION 13. Does Glauber dynamics for proper colorings mix in time order $n \log n$ if the number of colors is bigger than $\Delta + 2$, where Δ bounds the graph degrees? This is known to be polynomial for $q > (11/6)n$—see the Notes to Chapter 14.

QUESTION 14 (Gaussian elimination chain). Consider the group of $n \times n$ upper triangular matrices with entries in $\mathbb{Z}_2$. Select k uniformly from $\{2, \ldots, n\}$ and add the k-th row to the $(k-1)$-st row. The last column of the resulting matrices form a copy of the East model chain. Hence the lower bound of order n^2 for the East model (Theorem 7.15) is also a lower bound for the Gaussian elimination chain. Diaconis (personal communication) informed us he has obtained an upper bound of order n^4. What is the correct exponent?

APPENDIX A

Background Material

While writing my book I had an argument with Feller. He asserted that everyone said "random variable" and I asserted that everyone said "chance variable." We obviously had to use the same name in our books, so we decided the issue by a stochastic procedure. That is, we tossed for it and he won.

—J. Doob, as quoted in Snell (1997).

A.1. Probability Spaces and Random Variables

Modern probability is based on measure theory. For a full account, the reader should consult one of the many textbooks on the subject, e.g. Billingsley (1995) or Durrett (2005). The majority of this book requires only probability on countable spaces, for which Feller (1968) remains the best reference. For the purpose of establishing notation and terminology we record a few definitions here.

Given a set Ω, a ***σ-algebra*** is a collection $\mathcal{F}$ of subsets satisfying

(i) $\Omega \in \mathcal{F}$,
(ii) if $A_1, A_2, \ldots$ are elements of $\mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$, and
(iii) if $A \in \mathcal{F}$, then $A^c := \Omega \setminus A \in \mathcal{F}$.

A ***probability space*** is a set Ω together with a σ-algebra of subsets, whose elements are called ***events***.

The following are important cases.

EXAMPLE A.1. If a probability space Ω is a countable set, the σ-algebra of events is usually taken to be the collection of all subsets of Ω.

EXAMPLE A.2. If Ω is $\mathbb{R}^d$, then the ***Borel σ-algebra*** is the smallest σ-algebra containing all open sets.

EXAMPLE A.3. When Ω is the sequence space S^∞ for a finite set S, a set of the form

$$A_1 \times A_2 \times \cdots \times A_n \times S \times S \cdots, \quad A_k \subset S \text{ for all } k = 1, \ldots, n,$$

is called a ***cylinder*** set. The set of events in S^∞ is the smallest σ-algebra containing the cylinder sets.

Given a probability space, a ***probability measure*** is a non-negative function $\mathbf{P}$ defined on events and satisfying the following:

(i) $\mathbf{P}(\Omega) = 1$,
(ii) for any sequence of events $B_1, B_2, \ldots$ which are disjoint, meaning $B_i \cap B_j = \varnothing$ for $i \neq j$,

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(B_i).$$

If Ω is a countable set, a ***probability distribution*** (or sometimes simply a ***probability***) on Ω is a function $p : \Omega \to [0,1]$ such that $\sum_{\xi \in \Omega} p(\xi) = 1$. We will abuse notation and write, for any subset $A \subset \Omega$,

$$p(A) = \sum_{\xi \in A} p(\xi).$$

The set function $A \mapsto p(A)$ is a probability measure.

A function $f : \Omega \to \mathbb{R}$ is called ***measurable*** if $f^{-1}(B)$ is an event for all open sets B. If $\Omega = D$ is an open subset of $\mathbb{R}^d$ and $f : D \to [0, \infty)$ is a measurable function satisfying $\int_D f(x)dx = 1$, then f is called a ***density function***. Given a density function, the set function defined for Borel sets B by

$$\mu_f(B) = \int_B f(x)dx$$

is a probability measure. (Here, the integral is the *Lebesgue* interval. It agrees with the usual Riemann integral wherever the Riemann integral is defined.)

Given a probability space, a ***random variable*** X is a measurable function defined on Ω. We write $\{X \in A\}$ as shorthand for the set $\{\xi \in \Omega : X(\xi) \in A\} = X^{-1}(A)$. The ***distribution*** of a random variable X is the probability measure μ_X on $\mathbb{R}$ defined for Borel set B by

$$\mu_X(B) := \mathbf{P}\{X \in B\} := \mathbf{P}(\{X \in B\}).$$

We call a random variable X ***discrete*** if there is a finite or countable set S, called the ***support of*** X, such that $\mu_X(S) = 1$. In this case, the function

$$p_X(a) = \mathbf{P}\{X = a\}$$

is a probability distribution on S.

A random variable X is called ***absolutely continuous*** if there is a density function f on $\mathbb{R}$ such that

$$\mu_X(A) = \int_A f(x)dx.$$

For a discrete random variable X, the ***expectation*** $\mathbf{E}(X)$ can be computed by the formula

$$\mathbf{E}(X) = \sum_{x \in \mathbb{R}} x\mathbf{P}\{X = x\}.$$

(Note that there are at most countably many non-zero summands.) For an absolutely continuous random variable X, the expectation is computed by the formula

$$\mathbf{E}(X) = \int_{\mathbb{R}} x f_X(x)dx.$$

If X is a random variable, $g : \mathbb{R} \to \mathbb{R}$ is a function, and $Y = g(X)$, then the expectation $\mathbf{E}(Y)$ can be computed via the formulas

$$\mathbf{E}(Y) = \begin{cases} \int g(x)f(x)dx & \text{if } X \text{ is continuous with density } f, \\ \sum_{x \in S} g(x)p_X(x) & \text{if } X \text{ is discrete with support } S. \end{cases}$$

The ***variance*** of a random variable X is defined by

$$\mathrm{Var}(X) = \mathbf{E}\left((X - \mathbf{E}(X))^2 \right).$$

Fix a probability space and probability measure $\mathbf{P}$. Two events, A and B, are ***independent*** if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. Events $A_1, A_2, \ldots$ are independent if for any $i_1, i_2, \ldots, i_r$,

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \cdots \mathbf{P}(A_{i_r}).$$

Random variables $X_1, X_2, \ldots$ are independent if for all Borel sets $B_1, B_2, \ldots$, the events $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots$ are independent.

PROPOSITION A.4. *If X and Y and independent random variables such that* $\text{Var}(X)$ *and* $\text{Var}(Y)$ *exists, then* $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

There are two fundamental inequalities.

PROPOSITION A.5 (Markov's Inequality). *For a non-negative random variable X,*

$$\mathbf{P}\{X > a\} \leq \frac{\mathbf{E}(X)}{a}.$$

PROPOSITION A.6 (Chebyshev's Inequality). *For a random variable X with finite expectation* $\mathbf{E}(X)$ *and finite variance* $\text{Var}(X)$*,*

$$\mathbf{P}\{|X - \mathbf{E}(X)| > a\} \leq \frac{\text{Var}(X)}{a^2}.$$

A sequence of random variables (X_t) ***converges in probability*** to a random variable X if

$$\lim_{t \to \infty} \mathbf{P}\{|X_t - X| > \varepsilon\} = 0, \tag{A.1}$$

for all ε. This is denoted by $X_t \xrightarrow{\text{pr}} X$.

THEOREM A.7 (Weak Law of Large Numbers). *If (X_t) is a sequence of independent random variables such that* $\mathbf{E}(X_t) = \mu$ *and* $\text{Var}(X_t) = \sigma^2$ *for all t, then*

$$\frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{\text{pr}} \mu \quad \textit{as } T \to \infty.$$

PROOF. By linearity of expectation, $\mathbf{E}(T^{-1}\sum_{t=1}^{T} X_t) = \mu$, and by independence, $\text{Var}(T^{-1}\sum_{t=1}^{T} X_t) = \sigma^2/T$. Applying Chebyshev's inequality,

$$\mathbf{P}\left\{\left|\frac{1}{T}\sum_{t=1}^{T} X_t - \mu\right| > \varepsilon\right\} \leq \frac{\sigma^2}{T\varepsilon^2}.$$

For every $\varepsilon > 0$ fixed, the right-hand side tends to zero as $T \to \infty$. ■

THEOREM A.8 (Strong Law of Large Numbers). *Let $Z_1, Z_2, \ldots$ be a sequence of random variables with* $\mathbf{E}(Z_s) = 0$ *for all s and*

$$\text{Var}(Z_{s+1} + \cdots + Z_{s+k}) \leq Ck$$

for all s and k. Then

$$\mathbf{P}\left\{\lim_{t \to \infty} \frac{1}{t}\sum_{s=0}^{t-1} Z_s = 0\right\} = 1. \tag{A.2}$$

PROOF. Let $A_t := t^{-1} \sum_{s=0}^{t-1} Z_s$. Then

$$\mathbf{E}(A_t^2) = \frac{\mathbf{E}\left[\left(\sum_{s=0}^{t-1} Z_s\right)^2\right]}{t^2} \leq \frac{C}{t}.$$

Thus, $\mathbf{E}\left(\sum_{m=1}^{\infty} A_{m^2}^2\right) < \infty$, which in particular implies that

$$\mathbf{P}\left\{\sum_{m=1}^{\infty} A_{m^2}^2 < \infty\right\} = 1 \quad \text{and} \quad \mathbf{P}\left\{\lim_{m\to\infty} A_{m^2} = 0\right\} = 1. \tag{A.3}$$

For a given t, let m_t be such that $m_t^2 \leq t < (m_t+1)^2$. Then

$$A_t = \frac{1}{t}\left(m_t^2 A_{m_t^2} + \sum_{s=m_t^2}^{t-1} Z_s\right). \tag{A.4}$$

Since $\lim_{t\to\infty} t^{-1} m_t^2 = 1$, by (A.3),

$$\mathbf{P}\left\{\lim_{t\to\infty} t^{-1} m_t^2 A_{m_t^2} = 0\right\} = 1. \tag{A.5}$$

Defining $B_t := t^{-1} \sum_{s=m_t^2}^{t-1} Z_s$,

$$\mathbf{E}(B_t^2) = \frac{\text{Var}\left(\sum_{s=m_t^2}^{t-1} Z_s\right)}{t^2} \leq \frac{2Cm_t}{t^2} \leq \frac{2C}{t^{3/2}}.$$

Thus $\mathbf{E}(\sum_{t=0}^{\infty} B_t^2) < \infty$, and

$$\mathbf{P}\left\{\lim_{t\to\infty} \frac{\sum_{s=m_t^2+1}^{t} Z_s}{t} = 0\right\} = 1. \tag{A.6}$$

Putting together (A.5) and (A.6), from (A.4) we conclude that (A.2) holds. ■

Another important result about sums of independent and identically distributed random variables is that their distributions are approximately normal:

THEOREM A.9 (Central Limit Theorem). *For each n, let $X_{n,1}, X_{n,2}, \ldots, X_{n,n}$ be independent random variables, each with the same distribution having expectation $\mu = \mathbf{E}(X_{n,1})$ and variance $\sigma^2 = \text{Var}(X_{n,1})$. Let $S_n = \sum_{i=1}^{n} X_{n,i}$. Then for all $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right\} = \Phi(x),$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.

A.1.1. Limits of expectations. We know from calculus that if (f_n) is a sequence of functions defined on an interval I, satisfying for every $x \in I$

$$\lim_{n\to\infty} f_n(x) = f(x),$$

then it is not necessarily the case that

$$\lim_{n\to\infty} \int_I f_n(x)dx = \int_I f(x)dx.$$

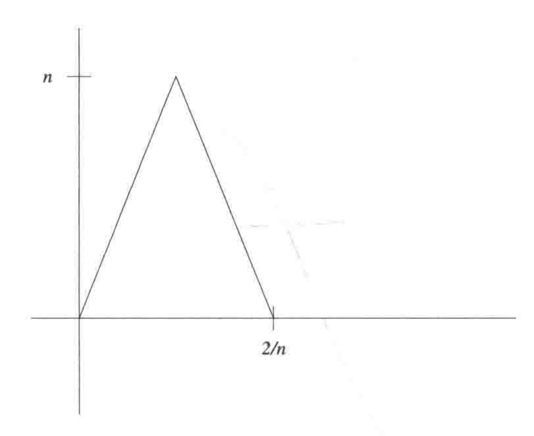


FIGURE A.1. A sequence of functions whose integrals do not converge to the integral of the limit.

As an example, consider the function g_n whose graph is shown in Figure A.1. The integral of this function is always 1, but for each $x \in [0,1]$, the limit $\lim_{n\to\infty} g_n(x) = 0$. That is,

$$\int_0^1 \lim_{n\to\infty} g_n(x)dx = 0 \neq 1 = \lim_{n\to\infty} \int_0^1 g_n(x)dx. \tag{A.7}$$

This example can be rephrased using random variables. Let U be a uniform random variable, and let $Y_n = g_n(U)$. Notice that $Y_n \to 0$. We have

$$\mathbf{E}(Y_n) = \mathbf{E}(g_n(U)) = \int g_n(x) f_U(x)dx = \int_0^1 g_n(x)dx,$$

as the density of U is $f_U = \mathbf{1}_{[0,1]}$. By (A.7),

$$\lim_{n\to\infty} \mathbf{E}(Y_n) \neq \mathbf{E}\left(\lim_{n\to\infty} Y_n\right).$$

Now that we have seen that we cannot always move a limit inside an expectation, can we ever? The answer is "yes", given some additional assumptions.

PROPOSITION A.10. *Let Y_n be a sequence of random variables and let Y be a random variable such that $\mathbf{P}\{\lim_{n\to\infty} Y_n = Y\} = 1$.*

(i) *If there is a constant K independent of n such that $|Y_n| < K$ for all n, then $\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y)$.*
(ii) *If there is a random variable Z such that $\mathbf{E}(|Z|) < \infty$ and $\mathbf{P}\{\,|Y_n| \leq |Z|\,\} = 1$ for all n, then $\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y)$.*
(iii) *If $\mathbf{P}\{Y_n \leq Y_{n+1}\} = 1$ for all n, then $\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y)$.*

Proposition A.10(i) is called the ***Bounded Convergence Theorem***, Proposition A.10(ii) is called the ***Dominated Convergence Theorem***, and Proposition A.10(iii) is called the ***Monotone Convergence Theorem***.

PROOF OF (I). For any $\varepsilon > 0$,

$$|Y_n - Y| \leq 2K\mathbf{1}_{\{|Y_n - Y| > \varepsilon/2\}} + \varepsilon/2,$$

and taking expectation above shows that

$$\begin{aligned}|\mathbf{E}(Y_n) - \mathbf{E}(Y)| &\leq \mathbf{E}\left(|Y_n - Y|\right) \\ &\leq 2K\mathbf{P}\left\{|Y_n - Y| > \varepsilon/2\right\} + \varepsilon/2.\end{aligned}$$

Since $\mathbf{P}\left\{|Y_n - Y| \geq \varepsilon/2\right\} \to 0$, by taking n sufficiently large,

$$|\mathbf{E}(Y_n) - \mathbf{E}(Y)| \leq \varepsilon.$$

That is, $\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y)$. ∎

For a proofs of (ii) and (iii), see Billingsley (1995).

A.2. Metric Spaces

A set M equipped with a function ρ measuring the distance between its elements is called a ***metric space***. In Euclidean space $\mathbb{R}^k$, the distance between vectors is measured by the norm $\|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$. On a graph, distance can be measured as the length of the shortest path connecting x and y. These are examples of metric spaces.

The function ρ must satisfy some properties to reasonably be called a distance. In particular, it should be symmetric, i.e., there should be no difference between measuring from a to b and measuring from b to a. Distance should never be negative, and there should be no two distinct elements which have distance zero. Finally, the distance $\rho(a, c)$ from a to c should never be greater than proceeding via a third point b and adding the distances $\rho(a, b) + \rho(b, c)$. For obvious reasons, this last property is called the ***triangle inequality***.

We summarize these properties here:

(i) $\rho(a, b) = \rho(b, a)$ for all $a, b \in M$.
(ii) $\rho(a, b) \geq 0$ for all $a, b \in M$, and $\rho(a, b) = 0$ only if $a = b$.
(iii) For any three elements $a, b, c \in M$,

$$\rho(a, c) \leq \rho(a, b) + \rho(b, c). \tag{A.8}$$

A.3. Linear Algebra

THEOREM A.11 (Spectral Theorem for Symmetric Matrices). *If M is a symmetric $m \times m$ matrix, then there exists a matrix U with $U^T U = I$ and a real diagonal matrix Λ such that $M = U^T \Lambda U$.*

(The matrix U^T is the ***transpose*** of U, whose entries are given by $U^T_{i,j} := U_{j,i}$.) A proof of Theorem A.11 can be found, for example, in Horn and Johnson (1990, Theorem 4.1.5).

Another way of formulating the Spectral Theorem is to say that there is an orthonormal basis of eigenvectors for M. The columns of U^T form one such basis, and the eigenvalue associated to the i-th column is $\lambda_i = \Lambda_{ii}$.

The variational characterization of the eigenvalues of a symmetric matrix is very useful:

THEOREM A.12 (Rayleigh-Ritz). *Let M be a symmetric matrix with eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

and associated eigenvectors $x_1, \ldots, x_n$. *Then*

$$\lambda_k = \max_{\substack{x \neq 0 \\ x \perp x_1, \ldots, x_{k-1}}} \frac{\langle x, Ax \rangle}{\langle x, x \rangle}.$$

See Horn and Johnson (1990, p. 178) for a discussion.

A.4. Miscellaneous

Stirling's formula says that

$$n! \sim \sqrt{2\pi} e^{-n} n^{n+1/2}, \tag{A.9}$$

where $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n b_n^{-1} = 1$.

More precise results are known, for example,

$$n! = \sqrt{2\pi} e^{-n} n^{n+1/2} e^{\varepsilon_n}, \qquad \frac{1}{12n+1} \leq \varepsilon_n \leq \frac{1}{12n}. \tag{A.10}$$

APPENDIX B

Introduction to Simulation

B.1. What Is Simulation?

Let X be a random unbiased bit:

$$\mathbf{P}\{X = 0\} = \mathbf{P}\{X = 1\} = \frac{1}{2}. \tag{B.1}$$

If we assign the value 0 to the "heads" side of a coin and the value 1 to the "tails" side, we can generate a bit which has the same distribution as X by tossing the coin.

Suppose now the bit is biased, so that

$$\mathbf{P}\{X = 1\} = \frac{1}{4}, \qquad \mathbf{P}\{X = 0\} = \frac{3}{4}. \tag{B.2}$$

Again using only our (fair) coin toss, we are able to easily generate a bit with this distribution: toss the coin twice and assign the value 1 to the result HH and the value 0 to the other three outcomes. Since the coin cannot remember the result of the first toss when it is tossed for the second time, the tosses are independent and the probability of two heads is 1/4. This recipe for generating observations of a random variable which has the same distribution (B.2) as X is called a ***simulation*** of X.

Consider the random variable U_n which is uniform on the finite set

$$\left\{0, \frac{1}{2^n}, \frac{2}{2^n}, \ldots, \frac{2^n - 1}{2^n}\right\}. \tag{B.3}$$

This random variable is a discrete approximation to the uniform distribution on $[0, 1]$. If our only resource is the humble fair coin, we are still able to simulate U_n: toss the coin n times to generate independent unbiased bits $X_1, X_2, \ldots, X_n$, and output the value

$$\sum_{i=1}^{n} \frac{X_i}{2^i}. \tag{B.4}$$

This random variable has the uniform distribution on the set in (B.3). (See Exercise B.1.)

Consequently, a sequence of independent and unbiased bits can be used to simulate a random variable whose distribution is close to uniform on $[0, 1]$. A sufficient number of bits should be used to ensure that the error in the approximation is small enough for any needed application. A computer can store a real number only to finite precision, so if the value of the simulated variable is to be placed in computer memory, it will be rounded to some finite decimal approximation. With this in mind, the discrete variable in (B.4) will be just as useful as a variable uniform on the interval of real numbers $[0, 1]$.

B.2. Von Neumann Unbiasing*

Suppose you have available an i.i.d. vector of *biased bits*, $X_1, X_2, \ldots, X_n$. That is, each X_k is a $\{0,1\}$-valued random variable, with $\mathbf{P}\{X_k = 1\} = p \neq 1/2$. Furthermore, suppose that we do not know the value of p. Can we convert this random vector into a (possibly shorter) random vector of independent and *unbiased* bits?

This problem was considered by von Neumann (1951) in his work on early computers. He described the following procedure: divide the original sequence of bits into pairs, discard pairs having the same value, and for each discordant pair 01 or 10, take the first bit. An example of this procedure is shown in Figure B.1; the extracted bits are shown in the second row.

original bits	00	11	01	01	10	00	10	10	11	10	01	···
extracted unbiased	·	·	0	0	1	·	1	1	·	1	0	···
discarded bits	0	1	·	·	·	0	·	·	1	·	·	···
XORed bits	0	0	1	1	1	0	1	1	0	1	1	···

(B.5)

FIGURE B.1. Extracting unbiased bits from biased bit stream.

Note that the number L of unbiased bits produced from $(X_1, \ldots, X_n)$ is itself a random variable. We denote by $(Y_1, \ldots, Y_L)$ the vector of extracted bits.

It is clear from symmetry that applying von Neumann's procedure to a bitstring $(X_1, \ldots, X_n)$ produces a bitstring $(Y_1, \ldots, Y_L)$ of random length L, which conditioned on $L = m$ is uniformly distributed on $\{0,1\}^m$. In particular, the bits of $(Y_1, \ldots, Y_L)$ are uniformly distributed and independent of each other.

How efficient is this method? For any algorithm for extracting random bits, let $N(n)$ be the number of fair bits generated using the first n of the original bits. The efficiency is measured by the asymptotic ***rate***

$$r(p) := \limsup_{n \to \infty} \frac{\mathbf{E}(N)}{n}. \tag{B.6}$$

Let $q := 1 - p$. For the von Neumann algorithm, each pair of bits has probability $2pq$ of contributing an extracted bit. Hence $\mathbf{E}(N(n)) = 2\left\lfloor \frac{n}{2} \right\rfloor pq$ and the efficiency is $r(p) = pq$.

The von Neumann algorithm throws out many of the original bits. These bits still contain some unexploited randomness. By converting the discarded 00's and 11's to 0's and 1's, we obtain a new vector $Z = (Z_1, Z_2, \ldots, Z_{\lfloor n/2 - L \rfloor})$ of bits. In the example shown in Figure B.1, these bits are shown on the third line.

Conditioned on $L = m$, the string $Y = (Y_1, \ldots, Y_L)$ and the string $Z = (Z_1, \ldots, Z_{\lfloor n/2-L \rfloor})$ are independent, and the bits $Z_1, \ldots, Z_{\lfloor n/2-L \rfloor}$ are independent of each other. The probability that $Z_i = 1$ is $p' = p^2/(p^2 + q^2)$. We can apply the von Neumann procedure again on the independent bits Z. Given that $L = m$, the expected number of fair bits we can extract from Z is

$$(\text{length of } Z)p'q' = \left\lfloor \frac{n}{2} - m \right\rfloor \left(\frac{p^2}{p^2+q^2} \right) \left(\frac{q^2}{p^2+q^2} \right). \tag{B.7}$$

Since $\mathbf{E}L = 2\left\lfloor \frac{n}{2} \right\rfloor pq$, the expected number of extracted bits is

$$(n + O(1))[(1/2) - pq]\left(\frac{p^2}{p^2+q^2}\right)\left(\frac{q^2}{p^2+q^2}\right). \tag{B.8}$$

Adding these bits to the original extracted bits yields a rate for the modified algorithm of

$$pq + [(1/2) - pq]\left(\frac{p^2}{p^2+q^2}\right)\left(\frac{q^2}{p^2+q^2}\right). \tag{B.9}$$

A third source of bits can be obtained by taking the XOR of adjacent pairs. (The XOR of two bits a and b is 0 if and only if $a = b$.) Call this sequence $U = (U_1, \ldots, U_{n/2})$. This is given on the fourth row in Figure B.1. It turns out that U is independent of Y and Z, and applying the algorithm on U yields independent and unbiased bits. It should be noted, however, that given $L = m$, the bits in U are not independent, as it contains exactly m 1's.

Note that when the von Neumann algorithm is applied to the sequence Z of discarded bits and to U, it creates a new sequence of discarded bits. The algorithm can be applied again to this sequence, improving the extraction rate.

Indeed, this can be continued indefinitely. This idea is developed in Peres (1992).

B.3. Simulating Discrete Distributions and Sampling

A Poisson random variable X with mean λ has mass function

$$p(k) := \frac{e^{-\lambda}\lambda^k}{k!}.$$

The variable X can be simulated using a uniform random variable U as follows: subdivide the unit interval into adjacent subintervals $I_1, I_2, \ldots$ where the length of I_k is $p(k)$. Because the chance that a random point in $[0,1]$ falls in I_k is $p(k)$, the index X for which $U \in I_X$ is a Poisson random variable with mean λ.

In principle, any discrete random variable can be simulated from a uniform random variable using this method. To be concrete, suppose X takes on the values $a_1, \ldots, a_N$ with probabilities $p_1, p_2, \ldots, p_N$. Let $F_k := \sum_{j=1}^{k} p_j$ (and $F_0 := 0$), and define $\varphi : [0,1] \to \{a_1, \ldots, a_N\}$ by

$$\varphi(u) := a_k \text{ if } F_{k-1} < u \leq F_k. \tag{B.10}$$

If $X = \varphi(U)$, where U is uniform on $[0,1]$, then $\mathbf{P}\{X = a_k\} = p_k$ (Exercise B.2).

One obstacle is that this recipe requires that the probabilities $(p_1, \ldots, p_N)$ are known exactly, while in many applications these are only known up to constant factor. This is a common situation, and many of the central examples treated in this book (such as the Ising model) fall into this category. It is common in applications to desire uniform samples from combinatorial sets whose sizes are not known.

Many problems are defined for a family of structures indexed by *instance size*. The efficiency of solutions is measured by the growth of the time required to run the algorithm as a function of instance size. If the run-time grows exponentially in instance size, the algorithm is considered impractical.

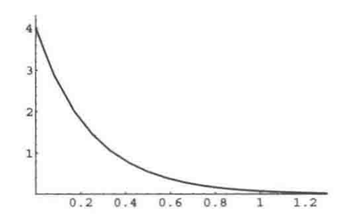


FIGURE B.2. $f(x) = 4e^{-4x}$, the exponential probability density function with rate 4.

B.4. Inverse Distribution Function Method

EXAMPLE B.1. Let U be a uniform random variable on $[0,1]$, and define $Y = -\lambda^{-1}\log(1-U)$. The distribution function of Y is

$$F(t) = \mathbf{P}\{Y \le t\} = \mathbf{P}\{-\lambda^{-1}\log(1-U) \le t\} = \mathbf{P}\{U \le 1 - e^{-\lambda t}\}. \qquad \text{(B.11)}$$

As U is uniform, the rightmost probability above equals $1 - e^{-\lambda t}$, the distribution function for an exponential random variable with rate λ. (The graph of an exponential density with $\lambda = 4$ is shown in Figure B.2.)

This calculation leads to the following algorithm:

(1) Generate U.
(2) Output $Y = -\lambda^{-1}\log(1-U)$.

The algorithm in Example B.1 is a special case of the ***inverse distribution function method*** for simulating a random variable with distribution function F, which is practical *provided that F can be inverted efficiently*. Unfortunately, there are not very many examples where this is the case.

Suppose that F is strictly increasing, so that its inverse function $F^{-1} : [0,1] \to \mathbb{R}$ is defined everywhere. Recall that F^{-1} is the function so that $F^{-1} \circ F(x) = x$ and $F \circ F^{-1}(y) = y$.

We now show how, using a uniform random variable U, to simulate X with distribution function F. For a uniform U, let $X = F^{-1}(U)$. Then

$$\mathbf{P}\{X \le t\} = \mathbf{P}\{F^{-1}(U) \le t\} = \mathbf{P}\{U \le F(t)\}. \qquad \text{(B.12)}$$

The last equality follows because F is strictly increasing, so $F^{-1}(U) \le t$ if and only if $F\left(F^{-1}(U)\right) \le F(t)$. Since U is uniform, the probability on the right can be easily evaluated to get

$$\mathbf{P}\{X \le t\} = F(t). \qquad \text{(B.13)}$$

That is, the distribution function of X is F.

B.5. Acceptance-Rejection Sampling

Suppose that we have a black box which on demand produces a uniform sample from a region R' in the plane, but what we really want is to sample from another region R which is contained in R' (see Figure B.3).

If independent points are generated, each uniformly distributed over R', until a point falls in R, then this point is a uniform sample from R (Exercise B.5).

Now we want to use this idea to simulate a random variable X with density function f given that we know how to simulate a random variable Y with density function g.

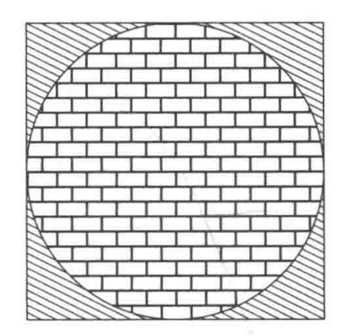


FIGURE B.3. R' is the diagonally hatched square, and R is the bricked circle.

We will suppose that

$$f(x) \leq Cg(x) \text{ for all } x, \tag{B.14}$$

for some constant C. We will see that good choices for the density g minimize the constant C. Because f and g both integrate to unity, $C \geq 1$.

Here is the algorithm:

(1) Generate a random variable Y having probability density function g.
(2) Generate a uniform random variable U.
(3) Conditional on $Y = y$, if $Cg(y)U \leq f(y)$, output the value y and halt.
(4) Repeat.

We now show that this method generates a random variable with probability density function f. Given that $Y = y$, the random variable $U_y := Cg(y)U$ is uniform on $[0, Cg(y)]$. By Exercise B.4, the point (Y, U_Y) is uniform over the region bounded between the graph of Cg and the horizontal axis. We halt the algorithm if and only if this point is also underneath the graph of f. By Exercise B.5, in this case, the point is uniformly distributed over the region under f. But again by Exercise B.4, the horizontal coordinate of this point has distribution f. (See Figure B.4.)

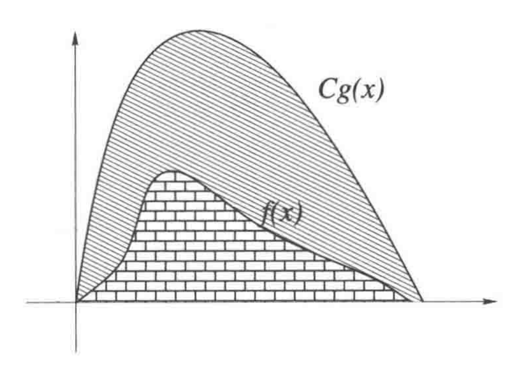


FIGURE B.4. The probability density function f lies below the scaled probability density function of g.

The value of C determines the efficiency of the algorithm. The probability that the algorithm terminates on any trial, given that $Y = y$, is $f(y)/Cg(y)$. Using the law of total probability, the unconditional probability is C^{-1}. The number of trials

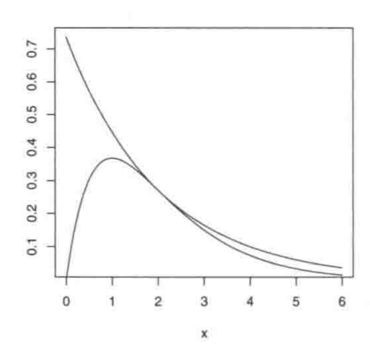


FIGURE B.5. The Gamma density for $\alpha = 2$ and $\lambda = 1$, along with $4e^{-1}$ times the exponential density of rate 1/2.

required is geometric, with success probability C^{-1}, and so the expected number of trials before terminating is C.

We comment here that there is a version of this method for discrete random variables; the reader should work on the details for herself.

EXAMPLE B.2. Consider the gamma distribution with parameters α and λ. Its probability density function is

$$f(x) = \frac{x^{\alpha-1}\lambda^{\alpha}e^{-\lambda x}}{\Gamma(\alpha)}. \tag{B.15}$$

(The function $\Gamma(\alpha)$ in the denominator is defined to normalize the density so that it integrates to unity. It has several interesting properties, most notably that $\Gamma(n) = (n-1)!$ for integers n.)

The distribution function does not have a nice closed-form expression, so inverting the distribution function does not provide an easy method of simulation.

We can use the rejection method here, when $\alpha > 1$, bounding the density by a multiple of the exponential density

$$g(x) = \mu e^{-\mu x}.$$

The constant C depends on μ, and

$$C = \sup_x \frac{[\Gamma(\alpha)]^{-1}(\lambda x)^{\alpha-1}\lambda e^{-\lambda x}}{\mu e^{-\mu x}}.$$

A bit of calculus shows that the supremum is attained at $x = (\alpha-1)/(\lambda-\mu)$ and

$$C = \frac{\lambda^{\alpha}(\alpha-1)^{\alpha-1}e^{1-\alpha}}{\Gamma(\alpha)\mu(\lambda-\mu)^{\alpha-1}}.$$

Some more calculus shows that the constant C is minimized for $\mu = \lambda/\alpha$, in which case

$$C = \frac{\alpha^{\alpha}e^{1-\alpha}}{\Gamma(\alpha)}.$$

The case of $\alpha = 2$ and $\lambda = 1$ is shown in Figure B.5, where $4e^{-1}\frac{1}{2}e^{-x/2}$ bounds the gamma density.

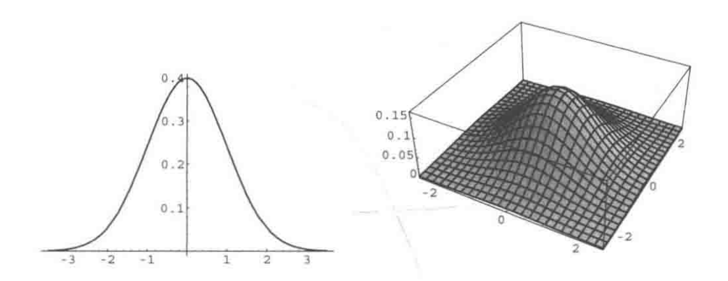


FIGURE B.6. The standard normal density on the left, and on the right the joint density of two independent standard normal variables.

We end the example by commenting that the exponential is easily simulated by the inverse distribution function method, as the inverse to $1-e^{-\mu x}$ is $(-1/\mu)\ln(1-u)$.

B.6. Simulating Normal Random Variables

Recall that a standard normal random variable has the "bell-shaped" probability density function specified by

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}. \tag{B.16}$$

The corresponding distribution function Φ is the integral

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}\,dt, \tag{B.17}$$

which cannot be evaluated in closed form. The inverse of Φ likewise cannot be expressed in terms of elementary functions. As a result the inverse distribution function method requires the numerical evaluation of Φ^{-1}. We present here another method of simulating from Φ which does not require the evaluation of the inverse of Φ.

Let X and Y be independent standard normal random variables. Geometrically, the ordered pair (X,Y) is a random point in the plane. The joint probability density function for (X,Y) is shown in Figure B.6.

We will write (R,Θ) for the representation of (X,Y) in polar coordinates and define $S := R^2 = X^2+Y^2$ to be the squared distance of (X,Y) to the origin.

The distribution function of S is

$$\mathbf{P}\{S \le t\} = \mathbf{P}\{X^2+Y^2 \le t\} = \iint_{D(\sqrt{t})} \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}\,dxdy, \tag{B.18}$$

where $D(\sqrt{t})$ is the disc of radius $\sqrt{t}$ centered at the origin. Changing to polar coordinates, this equals

$$\int_0^{\sqrt{t}}\int_0^{2\pi} \frac{1}{2\pi}e^{-\frac{r^2}{2}}\,rdrd\theta = 1-e^{-t/2}. \tag{B.19}$$

We conclude that S has an exponential distribution with mean 2.

To summarize, the squared radial part of (X,Y) has an exponential distribution, its angle has a uniform distribution, and these are independent.

Our standing assumption is that we have available independent uniform variables; here we need two, U_1 and U_2. Define $\Theta := 2\pi U_1$ and $S := -2\log(1 - U_2)$, so that Θ is uniform on $[0, 2\pi]$ and S is independent of Θ and has an exponential distribution.

Now let (X, Y) be the Cartesian coordinates of the point with polar representation $(\sqrt{S}, \Theta)$. Our discussion shows that X and Y are independent standard normal variables.

B.7. Sampling from the Simplex

Let Δ_n be the $n-1$-dimensional simplex:

$$\Delta_n := \left\{ (x_1, \ldots, x_n) \,:\, x_i \geq 0, \sum_{i=1}^{n} x_i = 1 \right\}. \tag{B.20}$$

This is the collection of probability vectors of length n. We consider here the problem of sampling from Δ_n.

Let $U_1, U_2, \ldots, U_{n-1}$ be i.i.d. uniform variables in $[0, 1]$, and define $U_{(k)}$ to be the k-th smallest among these.

Let $T : \mathbb{R}^{n-1} \to \mathbb{R}^n$ be the linear transformation defined by

$$T(u_1, \ldots, u_{n-1}) = (u_1, u_2 - u_1, \ldots, u_{n-1} - u_{n-2}, 1 - u_{n-1}).$$

Note that T maps the set $A_{n-1} = \{(u_1, \ldots, u_{n-1}) \,:\, u_1 \leq u_2 \leq \cdots \leq u_{n-1} \leq 1\}$ linearly to Δ_n, so Exercise B.8 and Exercise B.9 together show that $(X_1, \ldots, X_n) = T(U_{(1)}, \ldots, U_{(n-1)})$ is uniformly distributed on Δ_n.

We can now easily generate a sample from Δ_n: throw down $n-1$ points uniformly in the unit interval, sort them along with the points 0 and 1, and take the vector of successive distances between the points.

The algorithm described above requires sorting n variables. This sorting can, however, be avoided. See Exercise B.10.

B.8. About Random Numbers

Because most computer languages provide a built-in capability for simulating random numbers chosen independently from the uniform density on the unit interval $[0, 1]$, we will assume throughout this book that there is a ready source of independent uniform-$[0, 1]$ random variables.

This assumption requires some further discussion, however. Since computers are finitary machines and can work with numbers of only finite precision, it is in fact impossible for a computer to generate a continuous random variable. Not to worry: a discrete random variable which is uniform on, for example, the set in (B.3) is a very good approximation to the uniform distribution on $[0, 1]$, at least when n is large.

A more serious issue is that computers do not produce truly random numbers at all. Instead, they use deterministic algorithms, called ***pseudorandom number generators***, to produce sequences of numbers that *appear* random. There are many tests which identify features which are unlikely to occur in a sequence of independent and identically distributed random variables. If a sequence produced by a pseudorandom number generator can pass a battery of these tests, it is considered an appropriate substitute for random numbers.

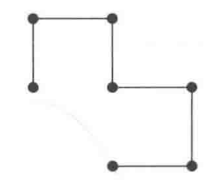


FIGURE B.7. A self-avoiding path

One technique for generating pseudorandom numbers is a ***linear congruential sequence*** (LCS). Let x_0 be an integer seed value. Given that x_{n-1} has been generated, let

$$x_n = (ax_{n-1} + b) \mod m. \tag{B.21}$$

Here a, b and m are fixed constants. Clearly, this produces integers in $\{0, 1, \ldots, m\}$; if a number in $[0, 1]$ is desired, divide by m.

The properties of $(x_0, x_1, x_2, \ldots)$ vary greatly depending on choices of a, b and m, and there is a great deal of art and science behind making judicious choices for the parameters. For example, if $a = 0$, the sequence does not look random at all!

Any linear congruential sequence is eventually periodic (Exercise B.12). The period of a LCS can be much smaller than m, the longest possible value.

The goal of any method for generating pseudorandom numbers is to generate output which is difficult to distinguish from truly random numbers using statistical methods. It is an interesting question whether a given pseudorandom number generator is good. We will not enter into this issue here, but the reader should be aware that the "random" numbers produced by today's computers are not in fact random, and sometimes this can lead to inaccurate simulations. For an excellent discussion of these issues, see Knuth (1997).

B.9. Sampling from Large Sets*

As discussed in Section 14.4, sampling from a finite set and estimating its size are related problems. Here we discuss the set of self-avoiding paths of length n and also mention domino tilings.

EXAMPLE B.3 (Self-avoiding walks). A self-avoiding walk in $\mathbb{Z}^2$ of length n is a sequence $(z_0, z_1, \ldots, z_n)$ such that $z_0 = (0, 0)$, $|z_i - z_{i-1}| = 1$, and $z_i \neq z_j$ for $i \neq j$. See Figure B.7 for an example of length 6. Let Ξ_n be the collection of all self-avoiding walks of length n. Chemical and physical structures such as molecules and polymers are often modeled as "random" self-avoiding walks, that is, as uniform samples from Ξ_n.

Unfortunately, no efficient algorithm for finding the size of Ξ_n is known. Nonetheless, we still desire (a practical) method for sampling uniformly from Ξ_n. We present a Markov chain in Example B.5 whose state space is the set of all self-avoiding walks of a given length and whose stationary distribution is uniform—but whose mixing time is not known.

EXAMPLE B.4 (Domino tilings). Domino tilings, sometimes also called ***dimer systems***, are another important family of examples for counting and sampling algorithms. A ***domino*** is a 2×1 or 1×2 rectangle, and, informally speaking, a ***domino tiling*** of a subregion of $\mathbb{Z}^2$ is a partition of the region into dominoes, disjoint except along their boundaries (see Figure B.8).

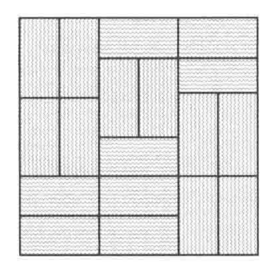


FIGURE B.8. A domino tiling of a 6×6 checkerboard.

Random domino tilings arise in statistical physics, and it was Kasteleyn (1961) who first computed that when n and m are both even, there are

$$2^{nm} \prod_{i=1}^{n/2} \prod_{j=1}^{m/2} \left(\cos^2 \frac{\pi i}{n+1} + \cos^2 \frac{\pi j}{m+1} \right)$$

domino tilings of an $n \times m$ grid.

The notion of a ***perfect matching*** (a set of disjoint edges together covering all vertices) generalizes domino tiling to arbitrary graphs, and much is known about counting and/or sampling perfect matchings on many families of graphs. See, for example, Luby, Randall, and Sinclair (1995) or Wilson (2004a). Section 22.2 discusses lozenge tilings, which correspond to perfect matchings on a hexagonal lattice.

EXAMPLE B.5 (Pivot chain for self-avoiding paths). The space Ξ_n of self-avoiding lattice paths of length n was described in Example B.3. These are paths in $\mathbb{Z}^2$ of length n which never intersect themselves.

Counting the number of self-avoiding paths is an unsolved problem. For more on this topic, see Madras and Slade (1993). Randall and Sinclair (2000) give an algorithm for approximately sampling from the uniform distribution on these walks.

We describe now a Markov chain on Ξ_n and show that it is irreducible. If the current state of the chain is the path $(0, v_1, \ldots, v_n) \in \Xi_n$, the next state is chosen by the following:

(1) Pick a value k from $\{0, 1, \ldots, n\}$ uniformly at random.
(2) Pick uniformly at random from the following transformations of $\mathbb{Z}^2$: rotations clockwise by $\pi/2$, π, $3\pi/2$, reflection across the x-axis, and reflection across the y-axis.
(3) Take the path from vertex k on, $(v_k, v_{k+1}, \ldots, v_n)$, and apply the transformation chosen in the previous step to this subpath only, taking v_k as the origin.
(4) If the resulting path is self-avoiding, this is the new state. If not, repeat.

An example move is shown in Figure B.9.

We now show that this chain is irreducible by proving that any self-avoiding path can be unwound to a straight line by a sequence of possible transitions. Since the four straight paths starting at $(0, 0)$ are rotations of each other and since any transition can also be undone by a dual transition, any self-avoiding path can be transformed into another. The proof below follows Madras and Slade (1993, Theorem 9.4.4).

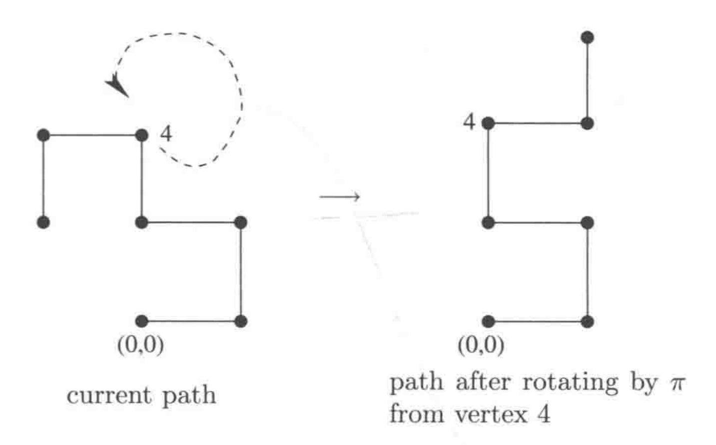


FIGURE B.9. Example of a single move of pivot chain for self-avoiding walk.

For a path $\xi \in \Xi_n$, put around ξ as small a rectangle as possible, and define $D = D(\xi)$ to be the sum of the length and the width of this rectangle. The left-hand diagram in Figure B.10 shows an example of this bounding rectangle. Define also $A = A(\xi)$ to be the number of interior vertices v of ξ where the two edges incident at v form an angle of π, that is, which look like either —•— or a vertical •. We first observe that $D(\xi) \leq n$ and $A(\xi) \leq n - 1$ for any $\xi \in \Xi_n$, and $D(\xi) + A(\xi) = 2n - 1$ if and only if ξ is a straight path. We show now that if ξ is any path different from the straight path, we can make a legal move—that is, a move having positive probability—to another path ξ' which has $D(\xi') + A(\xi') > D(\xi) + A(\xi)$.

There are two cases which we will consider separately.

Case 1. Suppose that at least one side of the bounding box does not contain either endpoint, 0 or v_n, of $\xi = (0, v_1, \ldots, v_n)$. This is the situation for the path on the left-hand side in Figure B.10. Let $k \geq 1$ be the smallest index so that v_k lies on this side. Obtain ξ' by taking ξ and reflecting its tail $(v_k, v_{k+1}, \ldots, v_n)$ across this box side. Figure B.10 shows an example of this transformation. The new path ξ' satisfies $D(\xi') > D(\xi)$ and $A(\xi') = A(\xi)$ (the reader should convince himself this is indeed true!)

Case 2. Suppose every side of the bounding box contains an endpoint of ξ. This implies that the endpoints are in opposing corners of the box. Let k be the largest index so that the edges incident to v_k form a right angle. The path ξ from v_k to v_n forms a straight line segment and must lie along the edge of the bounding box. Obtain ξ' from ξ by rotating this straight portion of ξ so that it lies outside the original bounding box. See Figure B.11.

This operation reduces one dimension of the bounding box by at most the length of the rotated segment, but increases the other dimension by this length. This shows that $D(\xi') \geq D(\xi)$. Also, we have strictly increased the number of straight angles, so $D(\xi') + A(\xi') > D(\xi) + A(\xi)$.

In either case, $D + A$ is strictly increased by the transformation, so continuing this procedure eventually leads to a straight line segment. This establishes that the pivot Markov chain is irreducible.

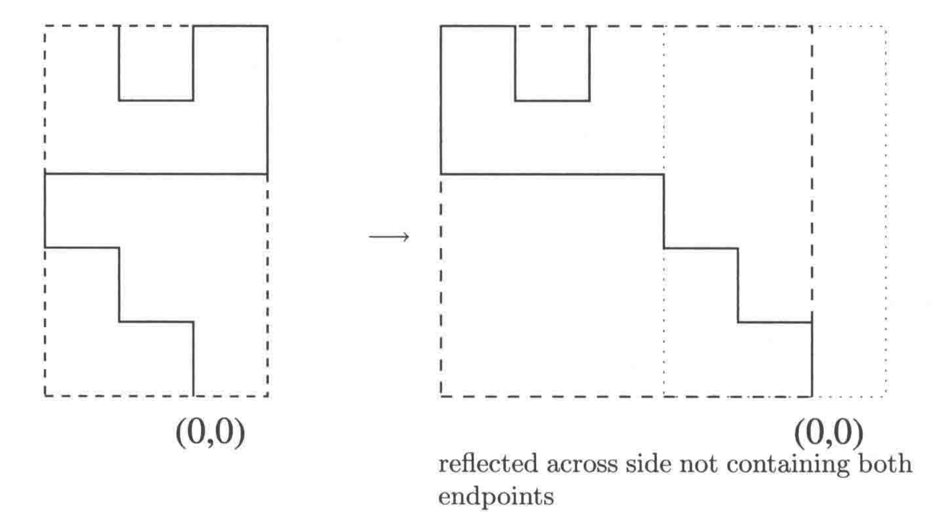


FIGURE B.10. A SAW without both endpoints in corners of bounding box.

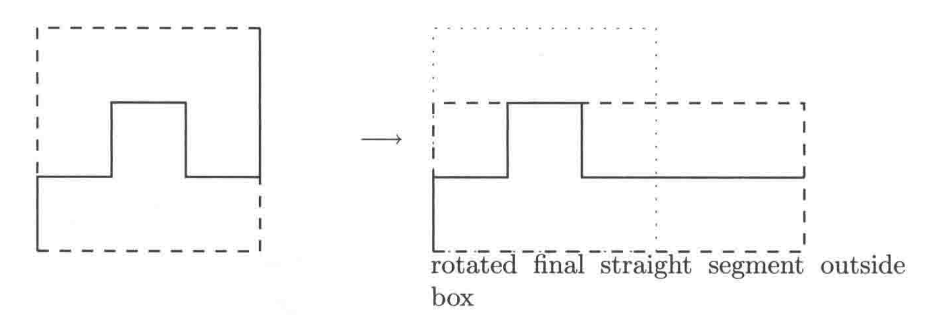


FIGURE B.11. A SAW with endpoints in opposing corners.

It is an open problem to analyze the convergence behavior of the pivot chain on self-avoiding walks. The algorithm of Randall and Sinclair (2000) uses a different underlying Markov chain to approximately sample from the uniform distribution on these walks.

Exercises

EXERCISE B.1. Check that the random variable in (B.4) has the uniform distribution on the set in (B.3).

EXERCISE B.2. Let U be uniform on $[0,1]$, and let X be the random variable $\varphi(U)$, where φ is defined as in (B.10). Show that X takes on the value a_k with probability p_k.

EXERCISE B.3. Describe how to use the inverse distribution function method to simulate from the probability density function

$$f(x) = \begin{cases} 2x & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

EXERCISE B.4. Show that if (Y, U_Y) is the pair generated in one round of the rejection sampling algorithm, then (Y, U_Y) is uniformly distributed over the region bounded between the graph of Cg and the horizontal axis. Conversely, if g is a density and a point is sampled from the region under the graph of g, then the projection of this point onto the x-axis has distribution g.

EXERCISE B.5. Let $R \subset R' \subset \mathbb{R}^k$. Show that if points uniform in R' are generated until a point falls in R, then this point is uniformly distributed over R. Recall that this means that the probability of falling in any subregion B of R is equal to $\mathrm{Vol}_k(B)/\mathrm{Vol}_k(R)$.

EXERCISE B.6. Argue that since the joint density $(2\pi)^{-1}\exp[-(x^2+y^2)/2]$ is a function of $s = x^2 + y^2$, the distribution of Θ must be uniform and independent of S.

EXERCISE B.7. Find a method for simulating the random variable Y with density

$$g(x) = e^{-|x|/2}.$$

Then use the rejection method to simulate a random variable X with the standard normal density given in (B.16).

EXERCISE B.8. Show that the vector $(U_{(1)}, \ldots, U_{(n-1)})$ is uniformly distributed over the set $A_{n-1} = \{(u_1, \ldots, u_{n-1}) : u_1 \le u_2 \le \cdots \le u_{n-1} \le 1\}$.

Let $T : \mathbb{R}^{n-1} \to \mathbb{R}^n$ be the linear transformation defined by

$$T(u_1, \ldots, u_{n-1}) = (u_1, u_2 - u_1, \ldots, u_{n-1} - u_{n-2}, 1 - u_{n-1}).$$

EXERCISE B.9. Suppose that X is uniformly distributed on a region A of $\mathbb{R}^d$, and the map $T : \mathbb{R}^d \to \mathbb{R}^r, d \le r$ is a linear transformation. A useful fact is that for a region $R \subset \mathbb{R}^d$,

$$\mathrm{Volume}_d(TR) = \sqrt{\det(T^tT)}\,\mathrm{Volume}(R),$$

where $\mathrm{Volume}_d(TR)$ is the d-dimensional volume of $TR \subset \mathbb{R}^r$. Use this to show that $Y = TX$ is uniformly distributed over TA.

EXERCISE B.10. (This exercise requires knowledge of the change-of-variables formula for d-dimensional random vectors.) Let $Y_1, \ldots, Y_n$ be i.i.d. exponential variables, and define

$$X_i = \frac{Y_i}{Y_1 + \cdots + Y_n}. \tag{B.22}$$

Show that $(X_1, \ldots, X_n)$ is uniformly distributed on Δ_n.

EXERCISE B.11. Let $U_1, U_2, \ldots, U_n$ be independent random variables, each uniform on the interval $[0, 1]$. Let $U_{(k)}$ be the k-th ***order statistic***, the k-th smallest among $\{U_1, \ldots, U_n\}$, so that

$$U_{(1)} < U_{(2)} < \cdots < U_{(n)}.$$

The purpose of this exercise is to give several different arguments that

$$\mathbf{E}\left(U_{(k)}\right) = \frac{k}{n+1}. \tag{B.23}$$

Fill in the details for the following proofs of (B.23):

(a) Find the density of $U_{(k)}$, and integrate.

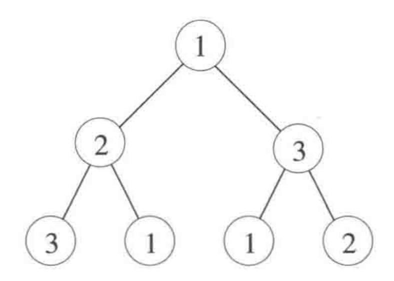


FIGURE B.12. A proper 3-coloring of a rooted tree. (As is common practice, we have placed the root at the top.)

(b) Find the density of $U_{(n)}$, and observe that given $U_{(n)}$, the other variables are the order statistics for uniforms on the interval $[0, U_{(n)}]$. Then apply induction.

(c) Let $Y_1, \ldots, Y_n$ be independent and identically distributed exponential variables with mean 1, and let $S_1 = Y_1, S_2 = Y_1 + Y_2, \ldots$ be their partial sums. Show that the random vector

$$\frac{1}{S_{n+1}} (S_1, S_2, \ldots, S_n) \tag{B.24}$$

has constant density on the simplex

$$\mathcal{A}_n = \{(x_1, \ldots, x_n) \; : \; 0 < x_1 < x_2 < \cdots < x_n < 1\}.$$

Conclude that (B.24) has the same law as the vector of order statistics.

EXERCISE B.12. Show that if $f : \{1, \ldots, m\} \to \{1, \ldots, m\}$ is any function and $x_n = f(x_{n-1})$ for all n, then there is an integer k such that $x_n = x_{n+k}$ eventually. That is, the sequence is eventually periodic.

EXERCISE B.13. Consider the following algorithm for sampling proper colorings on a rooted tree (see Figure B.12): choose the color of the root uniformly at random from $\{1, \ldots, q\}$. Given that colors have been assigned to all vertices up to depth d, for a vertex at depth $d+1$, assign a color chosen uniformly at random from

$$\{1, 2, \ldots, q\} \setminus \{\text{color of parent}\}. \tag{B.25}$$

(a) Verify that the coloring generated is uniformly distributed over all proper colorings.

(b) Similarly extend the sampling algorithms of Exercises 14.5 and 14.6 to the case where the base graph is an arbitrary rooted tree.

EXERCISE B.14. A nearest-neighbor path $0 = v_0, \ldots, v_n$ is ***non-reversing*** if $v_k \neq v_{k-2}$ for $k = 2, \ldots, n$. It is simple to generate a non-reversing path recursively. First choose v_1 uniformly at random from $\{(0,1), (1,0), (0,-1), (-1,0)\}$. Given that $v_0, \ldots, v_{k-1}$ is a non-reversing path, choose v_k uniformly from the three sites in $\mathbb{Z}^2$ at distance 1 from v_{k-1} but different from v_{k-2}.

Let Ξ_n^{nr} be the set of non-reversing nearest-neighbor paths of length n. Show that the above procedure generates a uniform random sample from Ξ_n^{nr}.

EXERCISE B.15. One way to generate a random self-avoiding path is to generate non-reversing paths until a self-avoiding path is obtained.

(a) Let $c_{n,4}$ be the number of paths in $\mathbb{Z}^2$ which do not contain loops of length 4 at indices $i \equiv 0 \mod 4$. More exactly, these are paths $(0,0) = v_0, v_1, \ldots, v_n$ so that $v_{4i} \neq v_{4(i-1)}$ for $i = 1, \ldots, n/4$. Show that

$$c_{n,4} \leq \left[4(3^3) - 8\right] \left[3^4 - 6\right]^{\lceil n/4 \rceil - 1} . \tag{B.26}$$

(b) Conclude that the probability that a random non-reversing path of length n is self-avoiding is bounded above by $e^{-\alpha n}$ for some fixed $\alpha > 0$.

Part (b) implies that if we try generating random non-reversing paths until we get a self-avoiding path, the expected number of trials required grows exponentially in the length of the paths.

Notes

On random numbers, von Neumann offers the following:

> "Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin" (von Neumann, 1951).

Iterating the von Neumann algorithm asymptotically achieves the optimal extraction rate of $-p\log_2 p - (1-p)\log_2(1-p)$, the entropy of a biased random bit (Peres, 1992). Earlier, a different optimal algorithm was given by Elias (1972), although the iterative algorithm has some computational advantages.

Further reading. For a stimulating and much wider discussion of univariate simulation techniques, Devroye (1986) is an excellent reference.

APPENDIX C

Solutions to Selected Exercises

Solutions to selected Chapter 1 exercises.

1.6. Fix x_0. Define for $k = 0, 1, \ldots, b-1$ the sets

$$\mathcal{C}_k := \{x \in \Omega : P^{mb+k}(x_0, x) > 0 \text{ for some } m\}. \tag{C.1}$$

Claim: Each x belongs to only one of the sets $\mathcal{C}_k$.

PROOF. Suppose $P^{mb+k}(x_0, x) > 0$ and $P^{m'b+j}(x_0, x) > 0$. Suppose, without loss of generality, that $j \le k$. There exists some r such that $P^r(x, x_0) > 0$, whence $r + mb + k \in \mathcal{T}(x_0)$. Therefore, b divides $r + k$. By the same reasoning, b divides $r + j$. Therefore, b must divide $r + k - (r + j) = k - j$. As $j \le k < b$, it must be that $k = j$. ■

Claim: The chain $(X_{bt})_{t=0}^{\infty}$, when started from $x \in \mathcal{C}_k$, is irreducible on $\mathcal{C}_k$.

PROOF. Let $x, y \in \mathcal{C}_k$. There exists r such that $P^r(x, x_0) > 0$. Also, by definition of $\mathcal{C}_k$, there exists m such that $P^{mb+k}(x_0, x) > 0$. Therefore, $r + mb + k \in \mathcal{T}(x_0)$, whence b divides $r + k$. Also, there exists m' such that $P^{m'b+k}(x_0, y) > 0$. Therefore, $P^{r+m'b+k}(x, y) > 0$. Since b divides $r + k$, we have $r + m'b + k = tb$ for some t. ■

Suppose that $x \in \mathcal{C}_i$ and $P(x, y) > 0$. By definition, there exists m such that $P^{mb+i}(x_0, y) > 0$. Since

$$P^{mb+i+1}(x_0, y) \ge P^{mb+i}(x_0, x) P(x, y) > 0,$$

it follows that $y \in \mathcal{C}_{i+1}$.

■

1.8. Observe that

$$\begin{aligned}
\pi(x) P^2(x, y) &= \pi(x) \sum_{z \in \Omega} P(x, z) P(z, y) \\
&= \sum_{z \in \Omega} \pi(z) P(z, x) P(z, y) \\
&= \sum_{z \in \Omega} \pi(z) P(z, y) P(z, x) \\
&= \sum_{z \in \Omega} \pi(y) P(y, z) P(z, x) \\
&= \pi(y) \sum_{z \in \Omega} P(y, z) P(z, x) \\
&= \pi(y) P^2(y, x).
\end{aligned}$$

Therefore, π is the stationary distribution for P^2. ■

1.11.

(a) Compute

$$\nu_n P(x) - \mu_n(x) = \frac{1}{n}\left(\mu P^n(x) - \mu(x)\right) \leq \frac{2}{n},$$

since any probability measure has weight at most 1 at x.

(b) Bolzano-Weierstrass, applied either directly in $\mathbb{R}^{|\Omega|}$ or iteratively: first take a subsequence that converges at x_1, then take a subsequence of that which converges at x_2, and so on. Either way, it's key that the weights of the measure are bounded and that the state space is finite.

(c) Part (a) gives stationarity, while the fact that the set of probability measures on Ω (viewed as a set in $\mathbb{R}^{|\Omega|}$) is closed gives that ν is a probability distribution.

■

Solutions to selected Chapter 2 exercises.

2.2. Let f_k be the expected value of the time until our gambler stops playing. Just as for the regular gambler's ruin, the values f_k are related:

$$f_0 = f_n = 0 \quad \text{and} \quad f_k = \frac{p}{2}(1 + f_{k-1}) + \frac{p}{2}(1 + f_{k+1}) + (1-p)(1 + f_k).$$

It is easy to check that setting $f_k = k(n-k)/p$ solves this system of equations. (Note that the answer is just what it should be. If she only bets a fraction p of the time, then it should take a factor of $1/p$ longer to reach her final state.) ■

2.3. Let (X_t) be a fair random walk on the set $\{-n, \ldots, n\}$, starting at the state 0 and absorbing at $\pm n$. By Proposition 2.1, the expected time for this walk to be absorbed is $(2n-n)(2n-n) = n^2$.

The walk described in the problem can be viewed as $n - |X_t|$. Hence its expected time to absorption is also n^2. ■

2.4.

$$\sum_{k=1}^{n} \frac{1}{k} \geq \sum_{k=1}^{n} \int_k^{k+1} \frac{dt}{t} = \int_1^{n+1} \frac{dt}{t} = \log(n+1) \geq \log n, \tag{C.2}$$

and

$$\sum_{k=1}^{n} \frac{1}{k} = 1 + \sum_{k=2}^{n} \frac{1}{k} \leq 1 + \sum_{k=2}^{n} \int_{k-1}^{k} \frac{dt}{t} = 1 + \int_1^n \frac{dt}{t} = 1 + \log n. \tag{C.3}$$

■

2.5.

$$\begin{aligned}
&\binom{d}{k+1} P(k+1,k) + \binom{d}{k-1} P(k-1,k) \\
&\quad = \frac{d!}{(k+1)!(d-k-1)!}\frac{k+1}{d} + \frac{d!}{(k-1)!(d-k+1)!}\frac{d-k+1}{d} \\
&\quad = \binom{d-1}{k} + \binom{d-1}{k-1} = \binom{d}{k}.
\end{aligned}$$

The last combinatorial identity, often called Pascal's identity, follows from splitting the set of k-element subsets of a d-element set into those which contain a distinguished element and those which do not. ■

2.8. Let φ be the function which maps $y \mapsto x$ and preserves P. Then

$$\hat{P}(z,w) = \frac{\pi(w)P(w,z)}{\pi(z)} = \frac{\pi(w)P(\varphi(w),\varphi(z))}{\pi(z)} = \hat{P}(w,z). \tag{C.4}$$

Note that the last equality follows since π is uniform, and so $\pi(x) = \pi(\varphi(x))$ for all x. ∎

2.10. Suppose that the reflected walk hits c at or before time n. It has probability at least $1/2$ of finishing at time n in $[c,\infty)$. (The probability can be larger than $1/2$ because of the reflecting at 0.) Thus

$$\mathbf{P}\left\{\max_{1\le j\le n} |S_j| \ge c\right\}\frac{1}{2} \le \mathbf{P}\left\{|S_n| \ge c\right\}.$$

∎

Solutions to selected Chapter 3 exercises.

3.1. Fix $x, y \in X$. Suppose first that $\pi(x)\Psi(x,y) \ge \pi(y)\Psi(y,x)$. In this case,

$$\pi(x)P(x,y) = \pi(x)\Psi(x,y)\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} = \pi(y)\Psi(y,x).$$

On the other hand, $\pi(y)P(y,x) = \pi(y)\Psi(y,x)$, so

$$\pi(x)P(x,y) = \pi(y)P(y,x). \tag{C.5}$$

Similarly, if $\pi(x)\Psi(x,y) < \pi(y)\Psi(x,y)$, then $\pi(x)P(x,y) = \pi(x)\Psi(x,y)$. Also,

$$\pi(y)P(y,x) = \pi(y)\Psi(y,x)\frac{\pi(x)\Psi(x,y)}{\pi(y)\Psi(y,x)} = \pi(x)\Psi(x,y).$$

Therefore, in this case, the detailed balance equation (C.5) is also satisfied. ∎

Solutions to selected Chapter 4 exercises.

4.1. By Proposition 4.2 and the triangle inequality we have

$$\begin{aligned}
\left\|\mu P^t - \pi\right\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{y\in\Omega} |\mu P^t(y) - \pi(y)| \\
&= \frac{1}{2}\sum_{y\in\Omega}\left|\sum_{x\in\Omega}\mu(x)P^t(x,y) - \sum_{x\in\Omega}\mu(x)\pi(y)\right| \\
&\le \frac{1}{2}\sum_{y\in\Omega}\sum_{x\in\Omega}\mu(x)|P^t(x,y) - \pi(y)| \\
&= \sum_{x\in\Omega}\mu(x)\frac{1}{2}\sum_{y\in\Omega}|P^t(x,y) - \pi(y)| \\
&= \sum_{x\in\Omega}\mu(x)\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \\
&\le \max_{x\in\Omega}\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}.
\end{aligned}$$

Since this holds for any μ, we have

$$\sup_{\mu}\left\|\mu P^t - \pi\right\|_{\mathrm{TV}} \le \max_{x\in\Omega}\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = d(t).$$

The opposite inequality holds, since the set of probabilities on Ω includes the point masses.

Similarly, if α and β are two probabilities on Ω, then

$$\begin{aligned}
\|\alpha P - \beta P\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{z\in\Omega}\left|\alpha P(z) - \sum_{w\in\Omega}\beta(w)P(w,z)\right| \\
&\le \frac{1}{2}\sum_{z\in\Omega}\sum_{w\in\Omega}\beta(w)|\alpha P(z) - P(w,z)| \\
&= \sum_{w\in\Omega}\beta(w)\frac{1}{2}\sum_{z\in\Omega}|\alpha P(z) - P(w,z)| \\
&= \sum_{w\in\Omega}\beta(w)\,\|\alpha P - P(w,\cdot)\|_{\mathrm{TV}} \\
&\le \max_{w\in\Omega}\|\alpha P - P(w,\cdot)\|_{\mathrm{TV}}\,. \qquad (C.6)
\end{aligned}$$

Thus, applying (C.6) with $\alpha = \mu$ and $\beta = \nu$ gives that

$$\|\mu P - \nu P\|_{\mathrm{TV}} \le \max_{y\in\Omega}\|\mu P - P(y,\cdot)\|_{\mathrm{TV}}\,. \qquad (C.7)$$

Applying (C.6) with $\alpha = \delta_y$, where $\delta_y(z) = \mathbf{1}_{\{z=y\}}$, and $\beta = \mu$ shows that

$$\|\mu P - P(y,\cdot)\|_{\mathrm{TV}} = \|P(y,\cdot) - \mu P\|_{\mathrm{TV}} \le \max_{x\in\Omega}\|P(y,\cdot) - P(x,\cdot)\|_{\mathrm{TV}}\,. \qquad (C.8)$$

Combining (C.7) with (C.8) shows that

$$\|\mu P - \nu P\|_{\mathrm{TV}} \le \max_{x,y\in\Omega}\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}}\,.$$

■

4.2. Define $A_n = n^{-1}\sum_{k=1}^{n} a_k$. Let $n_k \le m < n_{k+1}$. Then

$$A_m = \frac{n_k}{m}A_{n_k} + \frac{\sum_{j=n_k+1}^{m} a_j}{m}.$$

Because $n_k/n_{k+1} \le n_k/m \le 1$, the ratio n_k/m tends to 1. Thus the first term tends to a. If $|a_j| \le B$, then the absolute value of the second term is bounded by

$$B\left(\frac{n_{k+1} - n_k}{n_k}\right) \to 0.$$

Thus $A_m \to a$. ■

4.3. This is a standard exercise in manipulation of sums and inequalities. Apply Proposition 4.2, expand the matrix multiplication, apply the triangle inequality,

switch order of summation, and apply Proposition 4.2 once more:

$$\begin{aligned}
\|\mu P - \nu P\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{x\in\Omega} |\mu P(x) - \nu P(x)| \\
&= \frac{1}{2}\sum_{x\in\Omega} \left| \sum_{y\in\Omega} \mu(y)P(y,x) - \sum_{y\in\Omega} \nu(y)P(y,x) \right| \\
&= \frac{1}{2}\sum_{x\in\Omega} \left| \sum_{y\in\Omega} P(y,x)\left[\mu(y) - \nu(y)\right] \right| \\
&\le \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\Omega} P(y,x)\,|\mu(y) - \nu(y)| \\
&= \frac{1}{2}\sum_{y\in\Omega} |\mu(y) - \nu(y)| \sum_{x\in\Omega} P(y,x) \\
&= \frac{1}{2}\sum_{y\in\Omega} |\mu(y) - \nu(y)| \\
&= \|\mu - \nu\|_{\mathrm{TV}}\,.
\end{aligned}$$

■

4.5. For $i = 1, \ldots, n$, let $(X^{(i)}, Y^{(i)})$ be the optimal coupling of μ_i and ν_i. Let

$$\begin{aligned}
\boldsymbol{X} &:= (X^{(1)}, \ldots, X^{(n)}), \\
\boldsymbol{Y} &:= (Y^{(1)}, \ldots, X^{(n)}).
\end{aligned}$$

Since the distribution of $\boldsymbol{X}$ is μ and the distribution of $\boldsymbol{Y}$ is ν, the pair $(\boldsymbol{X}, \boldsymbol{Y})$ is a coupling of μ and ν. Thus

$$\|\mu - \nu\|_{\mathrm{TV}} \le \mathbf{P}\{\boldsymbol{X} \neq \boldsymbol{Y}\} \le \sum_{i=1}^{n} \mathbf{P}\{X_i = Y_i\} = \sum_{i=1}^{n} \|\mu_i - \nu_i\|_{\mathrm{TV}}\,.$$

■

Solutions to selected Chapter 5 exercises.

5.1. Consider the following coupling of the chain started from x and the chain started from π: run the chains independently until the time τ when they meet, and then run them together. Recall that by aperiodicity and irreducibility, there is some r so that $\alpha := \min_{x,y} P^r(x,y) \ge 0$.

Fix some state x_0. Then the probability that both chains, starting from say x and y, are not at x_0 after r steps is at most $(1-\alpha)$. If the two chains are not at x_0 after these r steps, the probability that they are not both at x_0 after another r steps is again $(1-\alpha)$. Continuing in this way, we get that $\mathbf{P}\{\tau > kr\} \le (1-\alpha)^k$. This shows that $\mathbf{P}\{\tau < \infty\} = 1$. ■

5.2. We show that

$$\mathbf{P}\{\tau_{\text{couple}} > kt_0\} \le (1-\alpha)^k, \tag{C.9}$$

from which the conclusion then follows by summing. An *unsuccessful coupling attempt* occurs at trial j if $X_t \neq Y_t$ for all $jt_0 < t \le (j+1)t_0$. Since (X_t, Y_t) is a Markovian coupling, so is (X_{t+jt_0}, Y_{t+jt_0}) for any j, and we can apply the

given bound on the probability of not coupling to any length-t_0 segment of the trajectories. Hence the probability of an unsuccessful coupling attempt at trial j is at most $(1-\alpha)$. It follows that the probability that all the first k attempts are unsuccessful is at most $(1-\alpha)^k$. ■

5.4. If τ_i is the coupling time of the i-th coordinate, we have seen already that $\mathbf{E}(\tau_i) \leq n^2/4$, so

$$\mathbf{P}\{\tau_i > dn^2\} \leq \frac{\mathbf{E}(\tau_i)}{kdn^2} \leq \frac{1}{4}.$$

Suppose that $\mathbf{P}\{\tau_i > (k-1)dn^2\} \leq 4^{-(k-1)}$. Then

$$\begin{aligned}\mathbf{P}\{\tau_i > kdn^2\} &= \mathbf{P}\{\tau_i > kdn^2 \mid \tau_i > (k-1)dn^2\}\mathbf{P}\{\tau_i > (k-1)dn^2\} \\ &\leq 4^{-1}4^{-(k-1)} \\ &= 4^{-k}.\end{aligned}$$

Letting $G_i = \{\tau_i > kdn^2\}$, we have $\mathbf{P}(G_i) \leq 4^{-1}$. Thus

$$\mathbf{P}\left\{\max_{1\leq i\leq d} \tau_i > kdn^2\right\} \leq \mathbf{P}\left(\bigcup_{i=1}^{d} G_i\right) \leq \sum_{i=1}^{d} \mathbf{P}(G_i) \leq d4^{-k}.$$

Taking $k = (1/2)\log_2(4d)$ makes the right-hand side equal $1/4$. Thus

$$t_{\text{mix}} \leq (1/2)[\log_2(4d)]dn^2 = O([d\log_2 d]n^2).$$

■

Solutions to selected Chapter 6 exercises.

6.1. Observe that if τ is a stopping time and r is a non-random and non-negative integer, then

$$\mathbf{1}_{\{\tau+r=t\}} = \mathbf{1}_{\{\tau=t-r\}} = f_{t-r}(X_0,\ldots,X_{t-r}),$$

where f_t is a function from Ω^{t+1} to $\{0,1\}$. Therefore, $\mathbf{1}_{\{\tau+r=t\}}$ is a function of $(X_0,\ldots,X_t)$, whence $\tau+r$ is a stopping time. ■

6.3. Let $\varepsilon := [2(2n-1)]^{-1}$. Let $\mu(v) = (2n-1)^{-1}$. For $v \neq v^\star$,

$$\begin{aligned}\sum_w \mu(w)P(w,v) &= \sum_{\substack{w\,:\,w\sim v\\ w\neq v}} \frac{1}{(2n-1)}\left[\frac{1}{2}-\varepsilon\right]\frac{1}{n-1} + \frac{1}{(2n-1)}\left[\frac{1}{2}+\varepsilon\right] \\ &= \frac{1}{(2n-1)}\left\{(n-1)\left[\frac{1}{2}-\varepsilon\right]\frac{1}{n-1} + \left[\frac{1}{2}+\varepsilon\right]\right\} \\ &= \frac{1}{2n-1}.\end{aligned}$$

Also,

$$\begin{aligned}\sum_w \mu(w)P(w,v^\star) &= (2n-2)\frac{1}{2n-1}\left[\frac{1}{2}-\varepsilon\right]\frac{1}{n-1} + \frac{1}{2n-1}\left(\frac{1}{2n-1}\right) \\ &= \frac{1}{2n-1}.\end{aligned}$$

■

6.5. By Exercise 6.4,

$$s(t) = s\left(t_0 \frac{t}{t_0}\right) \le s(t_0)^{t/t_0}.$$

Since $s(t_0) \le \varepsilon$ by hypothesis, applying Lemma 6.13 finishes the solution. ■

6.6. By the Monotone Convergence Theorem,

$$\mathbf{E}\left(\sum_{t=1}^{\tau} |Y_t|\right) = \sum_{t=1}^{\infty} \mathbf{E}\left(|Y_t|\mathbf{1}_{\{\tau \ge t\}}\right). \tag{C.10}$$

Since the event $\{\tau \ge t\}$ is by assumption independent of Y_t and $\mathbf{E}|Y_t| = \mathbf{E}|Y_1|$ for all $t \ge 1$, the right-hand side equals

$$\sum_{t=1}^{\infty} \mathbf{E}|Y_1|\mathbf{P}\{\tau \ge t\} = \mathbf{E}|Y_1| \sum_{t=1}^{\infty} \mathbf{P}\{\tau \ge t\} = \mathbf{E}|Y_1|\mathbf{E}(\tau) < \infty. \tag{C.11}$$

By the Dominated Convergence Theorem, since

$$\left| \sum_{t=1}^{\infty} Y_t \mathbf{1}_{\{\tau \ge t\}} \right| \le \sum_{t=1}^{\infty} |Y_t| \mathbf{1}_{\{\tau \ge t\}}$$

and (C.11) shows that the expectation of the non-negative random variable on the right-hand side above is finite,

$$\mathbf{E}\left(\sum_{t=1}^{\infty} Y_t \mathbf{1}_{\{\tau \ge t\}}\right) = \sum_{t=1}^{\infty} \mathbf{E}(Y_t \mathbf{1}_{\{\tau \ge t\}}) = \mathbf{E}(Y_1) \sum_{t=1}^{\infty} \mathbf{P}\{\tau \ge t\} = \mathbf{E}(Y_1)\mathbf{E}(\tau).$$

Now suppose that τ is a stopping time. For each t,

$$\{\tau \ge t\} = \{\tau \le t-1\}^c = \{(Y_1, \ldots, Y_{t-1}) \in B_{t-1}\}^c, \tag{C.12}$$

for some $B_{t-1} \subset \Omega^{t-1}$. Since the sequence (Y_t) is i.i.d., (C.12) shows that $\{\tau \ge t\}$ is independent of Y_t. ■

6.7. Let A be the set of vertices in one of the complete graphs making up G. Clearly, $\pi(A) = n/(2n-1) \ge 2^{-1}$.

On the other hand, for $x \notin A$,

$$P^t(x, A) = 1 - (1 - \alpha_n)^t \tag{C.13}$$

where

$$\alpha_n = \frac{1}{2}\left[1 - \frac{1}{2(n-1)}\right] \frac{1}{n-1} = \frac{1}{2n}\left[1 + o(1)\right].$$

The total variation distance can be bounded below:

$$\left\|P^t(x,\cdot) - \pi\right\|_{\text{TV}} \ge \pi(A) - P^t(x, A) \ge (1 - \alpha_n)^t - \frac{1}{2}. \tag{C.14}$$

Since

$$\log(1 - \alpha_n)^t \ge t(-\alpha_n - \alpha_n^2/2)$$

and $-1/4 \ge \log(3/4)$, if $t < \left[4\alpha_n(1 - \alpha_n/2)\right]^{-1}$, then

$$(1 - \alpha_n)^t - \frac{1}{2} \ge \frac{1}{4}.$$

This implies that $t_{\text{mix}}(1/4) \ge \frac{n}{2}\left[1 + o(1)\right]$. ■

6.9. Let τ be the first time all the vertices have been visited at least once, and let τ_k be the first time that vertex k has been reached. We have

$$\begin{aligned}
\mathbf{P}_0\{X_\tau = k\} &= \mathbf{P}_0\{X_\tau = k \mid \tau_{k-1} < \tau_{k+1}\}\mathbf{P}_0\{\tau_{k-1} < \tau_{k+1}\} \\
&\quad + \mathbf{P}_0\{X_\tau = k \mid \tau_{k+1} < \tau_{k-1}\}\mathbf{P}_0\{\tau_{k+1} < \tau_{k-1}\} \\
&= \mathbf{P}_{k-1}\{\tau_{k+1} < \tau_k\}\mathbf{P}_0\{\tau_{k-1} < \tau_{k+1}\} \\
&\quad + \mathbf{P}_{k+1}\{\tau_{k-1} < \tau_k\}\mathbf{P}_0\{\tau_{k+1} < \tau_{k-1}\} \\
&= \frac{1}{n-1}\mathbf{P}_0\{\tau_{k-1} < \tau_{k+1}\} + \frac{1}{n-1}\mathbf{P}_0\{\tau_{k+1} < \tau_{k-1}\} \\
&= \frac{1}{n-1}.
\end{aligned}$$

The identity $\mathbf{P}_{k+1}\{\tau_{k-1} < \tau_k\} = 1/(n-1)$ comes from breaking the cycle at k and using the gambler's ruin on the resulting segment. ■

Solutions to Chapter 7 exercises.

7.1. Let $Y_t^i = 2X_t^i - 1$. Since covariance is bilinear, $\text{Cov}(Y_t^i, Y_t^j) = 4\,\text{Cov}(X_t^i, X_t^j)$ and it is enough to check that $\text{Cov}(Y_t^i, Y_t^j) \le 0$.

If the i-th coordinate is chosen in the first t steps, the conditional expectation of Y_t^i is 0. Thus

$$\mathbf{E}(Y_t^i) = \left(1 - \frac{1}{n}\right)^t.$$

Similarly,

$$\mathbf{E}(Y_t^i Y_t^j) = \left(1 - \frac{2}{n}\right)^t$$

since we only have a positive contribution if both the coordinates i, j were not chosen in the first t steps. Finally,

$$\begin{aligned}
\text{Cov}\left(Y_t^i, Y_t^j\right) &= \mathbf{E}\left(Y_t^i Y_t^j\right) - \mathbf{E}\left(Y_t^i\right)\mathbf{E}\left(Y_t^j\right) \\
&= \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{1}{n}\right)^{2t} \\
&\le 0,
\end{aligned}$$

because $(1 - 2/n) < (1 - 1/n)^2$.

The variance of the sum $W_t = \sum_{i=1}^n X_t^i$ is

$$\text{Var}(N_t) = \sum_{i=1}^n \text{Var}(X_t^i) + \sum_{i \ne j} \text{Cov}(X_t^i, X_t^j) \le \sum_{i=1}^n \frac{1}{4}.$$

■

7.2.

$$\begin{aligned}
Q(S, S^c) &= \sum_{x \in S} \sum_{y \in S^c} \pi(x) P(x, y) \\
&= \sum_{y \in S^c} \left[\sum_{x \in \Omega} \pi(x) P(x, y) - \sum_{x \in S^c} \pi(x) P(x, y) \right] \\
&= \sum_{y \in S^c} \sum_{x \in \Omega} \pi(x) P(x, y) - \sum_{x \in S^c} \pi(x) \sum_{y \in S^c} P(x, y) \\
&= \sum_{y \in S^c} \pi(y) - \sum_{x \in S^c} \pi(x) \left[1 - \sum_{y \in S} P(x, y) \right] \\
&= \sum_{y \in S^c} \pi(y) - \sum_{x \in S^c} \pi(x) + \sum_{x \in S^c} \sum_{y \in S} \pi(x) P(x, y) \\
&= \sum_{x \in S^c} \sum_{y \in S} \pi(x) P(x, y) \\
&= Q(S^c, S).
\end{aligned}$$

■

7.3. Let $\{v_1, \ldots, v_n\}$ be the vertex set of the graph, and let (X_t) be the Markov chain started with the initial configuration $\boldsymbol{q}$ in which every vertex has color q.

Let $N : \Omega \to \{0, 1, \ldots, n\}$ be the number of sites in the configuration x colored with q. That is,

$$N(x) = \sum_{i=1}^{n} \mathbf{1}_{\{x(v_i) = q\}}. \tag{C.15}$$

We write N_t for $N(X_t)$.

We compare the mean and variance of the random variable N under the uniform measure π and under the measure $P^t(\boldsymbol{q}, \cdot)$. (Note that the distribution of $N(X_t)$ equals the distribution of N under $P^t(\boldsymbol{q}, \cdot)$.)

The distribution of N under the stationary measure π is binomial with parameters n and $1/q$, implying

$$E_\pi(N) = \frac{n}{q}, \quad \mathrm{Var}_\pi(N) = n\frac{1}{q}\left(1 - \frac{1}{q}\right) \leq \frac{n}{4}.$$

Let $X_i(t) = \mathbf{1}_{\{X_t(v_i) = q\}}$, the indicator that vertex v_i has color q. Since $X_i(t) = 0$ if and only if vertex v_i has been updated at least once by time t and the latest of these updates is *not* to color q, we have

$$\mathbf{E}_{\boldsymbol{q}}(X_i(t)) = 1 - \left[1 - \left(1 - \frac{1}{n}\right)^t\right] \frac{q-1}{q} = \frac{1}{q} + \frac{q-1}{q}\left(1 - \frac{1}{n}\right)^t$$

and

$$\mathbf{E}_{\boldsymbol{q}}(N_t) = \frac{n}{q} + \frac{n(q-1)}{q}\left(1 - \frac{1}{n}\right)^t.$$

Consequently,

$$\mathbf{E}_{\boldsymbol{q}}(N_t) - E_\pi(N) = \left(\frac{q-1}{q}\right) n \left(1 - \frac{1}{n}\right)^t.$$

The random variables $\{X_i(t)\}$ are negatively correlated; check that $Y_i = qX_i-(q-1)$ are negatively correlated as in the solution to Exercise 7.1. Thus,

$$\sigma^2 := \max\{\mathrm{Var}_{\boldsymbol{q}}(N_t), \mathrm{Var}_\pi(N)\} \leq \frac{n}{4},$$

and

$$|E_\pi(N) - \mathbf{E}_{\boldsymbol{q}}(N(X_t))| = \frac{n}{2}\left(1-\frac{1}{n}\right)^t \geq \sigma\frac{2(q-1)}{q}\sqrt{n}\left(1-\frac{1}{n}\right)^t.$$

Letting $r(t) = [2(q-1)/q]\sqrt{n}(1-n^{-1})^t$,

$$\begin{aligned}\log(r^2(t)) &= 2t\log(1-n^{-1}) + \frac{2(q-1)}{q}\log n \\ &\geq 2t\left(-\frac{1}{n}-\frac{1}{2n^2}\right) + \frac{2(q-1)}{q}\log n,\end{aligned} \tag{C.16}$$

where the inequality follows from $\log(1-x) \geq -x - x^2/2$, for $x \geq 0$. As in the proof of Proposition 7.13, it is possible to find a $c(q)$ so that for $t \leq (1/2)n\log n - c(q)n$, the inequality $r^2(t) \geq 32/3$ holds. By Proposition 7.8, $t_{\mathrm{mix}} \geq (1/2)n\log n - c(q)n$. ■

Solutions to selected Chapter 8 exercises.

8.1. Given a specific permutation $\eta \in \mathcal{S}_n$, the probability that $\sigma_k(j) = \eta(j)$ for $j = 1, 2, \ldots, k$ is equal to $\prod_{i=0}^{k-1}(n-i)^{-1}$, as can be seen by induction on $k = 1, \ldots, n-1$. ■

8.3.

(a) This is by now a standard application of the parity of permutations. Note that any sequence of moves in which the empty space ends up in the lower right corner must be of even length. Since every move is a single transposition, the permutation of the tiles (including the empty space as a tile) in any such position must be even. However, the desired permutation (switching two adjacent tiles in the bottom row) is odd.

(b) In fact, all even permutations of tiles can be achieved, but it is not entirely trivial to demonstrate. See Archer (1999) for an elementary proof and some historical discussion. Zhentao Lee discovered a new and elegant elementary proof during our 2006 MSRI workshop.

■

8.4. The function σ is a permutation if all of the images are distinct, which occurs with probability

$$p_n := \frac{n!}{n^n}.$$

By Stirling's formula, the expected number of trials needed is asymptotic to

$$\frac{e^n}{\sqrt{2\pi n}},$$

since the number of trials needed is geometric with parameter p_n. ■

8.5. The proposed method clearly yields a uniform permutation when $n = 1$ or $n = 2$. However, it fails to do so for for all larger values of n. One way to see this is to note that at each stage in the algorithm, there are n options. Hence the probability of each possible permutation must be an integral multiple of $1/n^n$. For $n \geq 3$, $n!$ is not a factor of n^n, so no permutation can have probability $1/n!$ of occurring. ■

8.6. False! Consider, for example, the distribution that assigns weight 1/2 each to the identity and to the permutation that lists the elements of $[n]$ in reverse order. ■

8.7. False! Consider, for example, the distribution that puts weight $1/n$ on all the cyclic shifts of a sorted deck: $123\dots n, 23\dots n1, \dots, n12\dots n-1$. ■

8.9. By Cauchy-Schwarz, for any permutation $\sigma \in \mathcal{S}_n$ we have

$$\varphi_\sigma = \sum_{k\in[n]} \varphi(k)\varphi(\sigma(k)) \leq \left(\sum_{k\in[n]} \varphi(k)^2\right)^{1/2} \left(\sum_{k\in[n]} \varphi(\sigma(k))^2\right)^{1/2} = \varphi_{\mathrm{id}}.$$

■

8.10. By the half-angle identity $\cos^2\theta = (\cos(2\theta)+1)/2$, we have

$$\sum_{k\in[n]} \cos^2\left(\frac{(2k-1)\pi}{2n}\right) = \frac{1}{2}\sum_{k\in[n]}\left(\cos\left(\frac{(2k-1)\pi}{n}\right)+1\right).$$

Now,

$$\sum_{k\in[n]} \cos\left(\frac{(2k-1)\pi}{n}\right) = \mathrm{Re}\left(e^{-\pi/n}\sum_{k\in[n]} e^{2k\pi/n}\right) = 0,$$

since the sum of the n-th roots of unity is 0. Hence

$$\sum_{k\in[n]} \cos^2\left(\frac{(2k-1)\pi}{2n}\right) = \frac{n}{2}.$$

■

8.11.

(a) Just as assigning t independent bits is the same as assigning a number chosen uniformly from $\{0,\dots,2^n-1\}$ (as we implicitly argued in the proof of Proposition 8.12), assigning a digit in base a and then a digit in base b, is the same as assigning a digit in base ab.

(b) To perform a forwards a-shuffle, divide the deck into a multinomially-distributed stacks, then uniformly choose an arrangement from all possible permutations that preserve the relative order within each stack. The resulting deck has at most a rising sequences, and there are a^n ways to divide and then riffle together (some of which can lead to identical permutations).

Given a permutation π with $r \leq a$ rising sequences, we need to count the number of ways it could possibly arise from a deck divided into a parts. Each rising sequence is a union of stacks, so the rising sequences together determine the positions of $r-1$ out of the $a-1$ dividers between stacks. The remaining $a-r$ dividers can be placed in any of the $n+1$ possible positions, repetition allowed, irrespective of the positions of the $r-1$ dividers already determined.

For example: set $a = 5$ and let $\pi \in \mathcal{S}_9$ be 152738946. The rising sequences are $(1,2,3,4)$, $(5,6)$, and $(7,8,9)$, so there must be packet divisions between 4 and 5 and between 6 and 7, and two additional dividers must be placed.

This is a standard choosing-with-repetition scenario. We can imagine building a row of length $n + (a - r)$ objects, of which n are numbers and $a - r$ are dividers. There are $\binom{n+a-r}{n}$ such rows.

Since each (division, riffle) pair has probability $1/a^n$, the probability that π arises from an a-shuffle is exactly $\binom{n+a-r}{n}/a^n$.

■

Solutions to selected Chapter 9 exercises.

9.1. Let $d \geq 2$. Let $U_{-d+1} = 1$, and let

$$U_{-d+2}, U_{-d+1}, \ldots, U_0, \ldots, U_n$$

be i.i.d. and uniform on $[0,1]$. Let $V_1, \ldots, V_d$ be the order statistics for $U_{-d+1}, \ldots, U_0$, let $V_0 = 0$, and define, for $1 \leq j \leq d$,

$$A_t^{(j)} := |\{-d+1 \leq j \leq t\} : V_{j-1} < U_j \leq V_j\}|.$$

Observe that $A_0^{(j)} = 1$ for all $1 \leq j \leq d$.

Consider an urn with initially d balls, each of a different color. At each unit of time, a ball is drawn at random and replaced along with an additional ball of the same color. Let $B_t^{(j)}$ be the number of balls of color j after t draws.

Claim: The distribution of $(\{A_t^{(j)}\}_{j=1}^d)$ and $(\{B_t^{(j)}\}_{j=1}^d)$ are the same.

PROOF OF CLAIM. Conditioned on the relative positions of $(U_{-d+2}, \ldots, U_t)$, the relative position of U_{t+1} is uniform on all $t + d$ possibilities. Thus the conditional probability that U_{t+1} falls between V_{j-1} and V_j is proportional to the number among $U_0, \ldots, U_t$ which fall in this interval, plus one. Thus, the conditional probability that $A_t^{(j)}$ increases by one equals $A_t^{(j)}/(t+d)$. This shows the transition probabilities for $\{A_t^{(j)}\}_{j=1}^d$ are exactly equal to those for $\{B_t^{(j)}\}_{j=1}^d$. Since they begin with the same initial distribution, their distributions are the same for $t = 0, \ldots, n$. ■

It is clear that the distribution of the d-dimensional vector $(A_t^{(1)}, \ldots, A_t^{(d)})$ is uniform over

$$\left\{(x_1, \ldots, x_d) : \sum_{i=1}^d x_i = t + d\right\}.$$

Construct a flow θ on the box $\{1, 2, \ldots, n\}^d$ as in the proof of Proposition 9.16 by defining for edges in the lower half of the box

$$\theta(e) = \mathbf{P}\{\text{Polya's } d\text{-colored process goes thru } e\}.$$

From above, we know that the process is equally likely to pass through each d-tuple $\boldsymbol{x}$ with $\sum x_i = k + d$. There are $\binom{k+d-1}{d-1}$ such d-tuples, whence each such edge has flow $[\binom{k+d-1}{d-1}]^{-1}$. There are constants c_1, c_2 (depending on d) such that $c_1 \leq \binom{k+d-1}{d-1}/k^{d-1} \leq c_2$. Therefore, the energy is bounded by

$$\mathcal{E}(\theta) \leq 2\sum_{k=1}^{n-1} \binom{k+d-1}{d-1}^{-2} \binom{k+d-1}{d-1} \leq c_3(d) \sum_{k=1}^{n-1} k^{-d+1} \leq c_4(d),$$

the last bound holding only when $d \geq 3$. ∎

9.5. In the new network obtained by gluing the two vertices, the voltage function cannot be the same as the voltage in the original network. Thus the corresponding current flow must differ. However, the old current flow remains a flow. By the uniqueness part of Thomson's Principle (Theorem 9.10), the effective resistance must change. ∎

9.8. Let W_1 be a voltage function for the unit current flow from x to y so that $W_1(x) = \mathcal{R}(x \leftrightarrow y)$ and $W_1(y) = 0$. Let W_2 be a voltage function for the unit current flow from y to z so that $W_2(y) = \mathcal{R}(y \leftrightarrow z)$ and $W_2(z) = 0$. By harmonicity (the maximum principle) at all vertices v we have

$$0 \leq W_1(v) \leq \mathcal{R}(x \leftrightarrow y) \tag{C.17}$$

$$0 \leq W_1(v) \leq \mathcal{R}(y \leftrightarrow z) \tag{C.18}$$

Recall the hint. Thus $W_3 = W_1 + W_2$ is a voltage function for the unit current flow from x to z and

$$\mathcal{R}(x \leftrightarrow z) = W_3(x) - W_3(z) = \mathcal{R}(x \leftrightarrow y) + W_2(x) - W_1(z). \tag{C.19}$$

Applying (C.18) gives $W_2(x) \leq \mathcal{R}(y \leftrightarrow z)$ and (C.17) gives $W_1(z) \geq 0$ so finally by (C.19) we get the triangle inequality. ∎

Solutions to selected Chapter 10 exercises.

10.4.

(a) By the Commute Time Identity (Proposition 10.6) and Example 9.7, the value is $2(n-1)(m-h)$.

(b) By (a), these pairs are clearly maximal over all those which are at the same level. If a is at level m and b is at level h, where $h < m$, let c be a descendant of b at level m. Since every walk from a to c must pass through b, we have $\mathbf{E}_a \tau_b \leq \mathbf{E}_a \tau_c$. A similar argument goes through when a is higher than b.

∎

10.5. Observe that $h_m(k)$ is the mean hitting time from k to 0 in G_k, which implies that $h_m(k)$ is monotone increasing in k. (This is intuitively clear but harder to prove directly on the cube.) The expected return time from o to itself in the hypercube equals 2^m but considering the first step, it also equals $1 + h_m(1)$. Thus

$$h_m(1) = 2^m - 1. \tag{C.20}$$

To compute $h_m(m)$, use symmetry and the Commute Time Identity. The effective resistance between 0 and m in G_m is $\mathcal{R}(0 \leftrightarrow m) = \sum_{k=1}^{m} [k\binom{m}{k}]^{-1}$. In this sum all but the first and last terms are negligible: the sum of the other terms is at most $4/m^2$ (check!). Thus

$$2h_m(m) = 2\mathcal{R}(0 \leftrightarrow m)|\text{edges}(G_m)| \leq 2\left(\frac{2}{m} + \frac{4}{m^2}\right)(m2^{m-1}),$$

so

$$h_m(m) \leq 2^m(1 + 2/m). \tag{C.21}$$

Equality (C.20) together with (C.21) and monotonicity concludes the proof. ∎

10.7. By Lemma 10.10,

$$\begin{aligned}2\mathbf{E}_a(\tau_{bca}) &= [\mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_c) + \mathbf{E}_c(\tau_a)] + [\mathbf{E}_a(\tau_c) + \mathbf{E}_c(\tau_b) + \mathbf{E}_b(\tau_a)] \\ &= [\mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_a)] + [\mathbf{E}_b(\tau_c) + \mathbf{E}_c(\tau_b)] + [\mathbf{E}_c(\tau_a) + \mathbf{E}_a(\tau_c)].\end{aligned}$$

Then the conclusion follows from Proposition 10.6. ■

10.8. Taking expectations in (10.32) yields

$$\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) = \mathbf{E}_x(\tau_z) + \mathbf{P}_x\{\tau_z < \tau_a\}\left[\mathbf{E}_z(\tau_a) + \mathbf{E}_a(\tau_z)\right],$$

which shows that

$$\mathbf{P}_x\{\tau_z < \tau_a\} = \frac{\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) - \mathbf{E}_x(\tau_z)}{\mathbf{E}_z(\tau_a) + \mathbf{E}_a(\tau_z)}, \tag{C.22}$$

without assuming reversibility.

In the reversible case, the cycle identity (Lemma 10.10) yields

$$\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) - \mathbf{E}_x(\tau_z) = \mathbf{E}_a(\tau_x) + \mathbf{E}_z(\tau_a) - \mathbf{E}_z(\tau_x). \tag{C.23}$$

Adding the two sides of (C.23) together establishes that

$$\begin{aligned}&\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) - \mathbf{E}_z(\tau_z) \\ &\quad = \frac{1}{2}\left\{[\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_x)] + [\mathbf{E}_a(\tau_z) + \mathbf{E}_z(\tau_a)] - [\mathbf{E}_x(\tau_z) + \mathbf{E}_z(\tau_x)]\right\}.\end{aligned}$$

Let $c_G = \sum_{x \in V} c(x) = 2\sum_e c(e)$, as usual. Then by the Commute Time Identity (Proposition 10.6), the denominator in (C.22) is $c_G \mathcal{R}(a \leftrightarrow z)$ and the numerator is $(1/2)c_G\left[\mathcal{R}(x \leftrightarrow a) + \mathcal{R}(a \leftrightarrow z) - \mathcal{R}(z \leftrightarrow x)\right]$. ■

10.9.

$$\begin{aligned}\sum_{k=0}^{\infty} c_k s^k &= \sum_{k=0}^{\infty}\sum_{j=0}^{k} a_j b_{k-j} s^k \\ &= \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} a_j s^j b_{k-j} s^{k-j} \mathbf{1}_{\{k \geq j\}} \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a^j s^j b_{k-j} s^{k-j} \mathbf{1}_{\{k \geq j\}} \\ &= \sum_{j=0}^{\infty} a^j s^j \sum_{k=0}^{\infty} b_{k-j} s^{k-j} \mathbf{1}_{\{k \geq j\}} \\ &= \sum_{j=0}^{\infty} a^j s^j \sum_{\ell=0}^{\infty} b_\ell s^\ell \\ &= A(s)B(s).\end{aligned}$$

The penultimate equality follows from letting $\ell = k - j$. The reader should check that the change of the order of summation is justified. ■

Solutions to selected Chapter 11 exercises.

11.1.

(a) Use the fact that, since the B_j's partition B, $\mathbf{E}(Y \mid B) = \sum_j \mathbf{P}(B_j)\mathbf{E}(Y \mid B_j)$.

(b) Many examples are possible; a small one is $\Omega = B = \{1,2,3\}$, $Y = \mathbf{1}_{\{1,3\}}$, $B_1 = \{1,2\}$, $B_2 = \{2,3\}$, $M = 1/2$.

■

Solutions to selected Chapter 12 exercises.

12.1.

(a) For any function f,

$$\|Pf\|_\infty = \max_{x\in\Omega}\left|\sum_{y\in\Omega} P(x,y)f(y)\right| \le \|f\|_\infty.$$

If $P\varphi = \lambda\varphi$, then $\|Pf\|_\infty = |\lambda|\,\|f\|_\infty \le \|f\|_\infty$. This implies that $|\lambda| \le 1$.

(c) Assume that a divides $\mathcal{T}(x)$. If b is the gcd of $\mathcal{T}(x)$, then a divides b. If ω is an a-th root of unity, then $\omega^b = 1$.

Let $\mathcal{C}_j$ be the subset of Ω defined in (C.1), for $j = 0, \ldots, b$. It is shown in the solution to Exercise 1.6 that

(i) there is a unique $j(x) \in \{0, \ldots, b-1\}$ such that $x \in \mathcal{C}_{j(x)}$ and

(ii) if $P(x,y) > 0$, then $j(y) = j(x) \oplus 1$. (Here $\oplus$ is addition modulo b.)

Let $f : \Omega \to \mathbb{C}$ be defined by $f(x) = \omega^{j(x)}$. We have that, for some $\ell \in \mathbb{Z}$,

$$Pf(x) = \sum_{y\in\Omega} P(x,y)\omega^{j(y)} = \omega^{j(x)\oplus 1} = \omega^{j(x)+1+\ell b} = \omega\omega^{j(x)} = \omega f(x).$$

Therefore, $f(x)$ is an eigenfunction with eigenvalue ω.

Let ω be an a-th root of unit, and suppose that $\omega f = Pf$ for some f. Choose x such that $|f(x)| = r := \max_{y\in\Omega}|f(y)|$. Since

$$\omega f(x) = Pf(x) = \sum_{y\in\Omega} P(x,y)f(y),$$

taking absolute values shows that

$$r \le \sum_{y\in\Omega} P(x,y)|f(y)| \le r.$$

We conclude that if $P(x,y) > 0$, then $|f(y)| = r$. By irreducibility, $|f(y)| = r$ for all $y \in \Omega$.

Since the average of complex numbers of norm r has norm r if and only if all the values have the same angle, it follows that $f(y)$ has the same value for all y with $P(x,y) > 0$. Therefore, if $P(x,y) > 0$, then $f(y) = \omega f(x)$. Now fix $x_0 \in \Omega$ and define for $j = 0, 1, \ldots, a-1$

$$\mathcal{C}_j = \{z \in \Omega : f(z) = \omega^j f(x_0)\}.$$

It is clear that if $P(x,y) > 0$ and $x \in \mathcal{C}_j$, then $x \in \mathcal{C}_{j\oplus 1}$, where $\oplus$ is addition modulo a. Also, it is clear that if $t \in \mathcal{T}(x_0)$, then a divides t.

■

12.3. Let f be an eigenfunction of P with eigenvalue μ. Then

$$\mu f = \tilde{P} f = \frac{Pf + f}{2}.$$

Rearranging shows that $(2\mu - 1)$ is an eigenvalue of P. Thus $2\mu - 1 \geq -1$, or equivalently, $\mu \geq 0$. ■

12.4. We first observe that

$$\begin{aligned} E_\pi(P^t f) &= \sum_{x\in\Omega} (P^t f)(x)\pi(x) = \sum_{x\in\Omega}\sum_{y\in\Omega} f(y)P^t(x,y)\pi(x) \\ &= \sum_{y\in\Omega} f(y) \sum_{x\in\Omega} \pi(x)P^t(x,y) = \sum_{y\in\Omega} f(y)\pi(y) = E_\pi(f). \end{aligned}$$

Since we take the first eigenfunction f_1 to be the constant function with value 1, we have $E_\pi(P^t f) = E_\pi(f) = \langle P^t f, f_1\rangle_\pi$. Therefore, it follows from (12.5) that $P^t f - E_\pi(P^t f) = \sum_{j=2}^{\Omega} \langle f, f_j\rangle_\pi f_j \lambda_j^t$. Since the f_j's are an orthonormal basis,

$$\begin{aligned} \mathrm{Var}_\pi(f) &= \langle P^t f - E_\pi(P^t f), P^t f - E_\pi(P^t f)\rangle_\pi \\ &= \sum_{j=2}^{|\Omega|} \langle f, f_j\rangle_\pi^2 \lambda_j^{2t} \\ &\leq (1-\gamma_\star)^2 \sum_{j=2}^{|\Omega|} \langle f, f_j\rangle_\pi^2. \end{aligned}$$

We observe that

$$\sum_{j=2}^{|\Omega|} \langle f, f_j\rangle_\pi^2 = \sum_{j=1}^{|\Omega|} \langle f, f_j\rangle_\pi^2 - E_\pi^2(f) = E_\pi(f^2) - E_\pi^2(f) = \mathrm{Var}_\pi(f).$$

■

12.6. According to (12.2),

$$\frac{P^{2t+2}(x,x)}{\pi(x)} = \sum_{j=1}^{|\Omega|} f_j(x)^2 \lambda_j^{2t+2}.$$

Since $\lambda_j^2 \leq 1$ for all j, the right-hand side is bounded above by $\sum_{j=1}^{|\Omega|} f_j(x)^2 \lambda_j^{2t}$, which equals $P^{2t}(x,x)/\pi(x)$. ■

12.7. A computation verifies the claim:

$$\begin{aligned} (P_1 \otimes P_2)(\varphi \otimes \psi)(x,y) &= \sum_{(z,w)\in\Omega_1\times\Omega_2} P_1(x,z)P_2(y,w)\varphi(z)\psi(w) \\ &= \sum_{z\in\Omega_1} P_1(x,z)\varphi(z) \sum_{w\in\Omega_2} P_2(y,w)\psi(w) \\ &= [P_1\varphi(x)]\,[P_2\psi(y)] \\ &= \lambda\mu\varphi(x)\psi(y) \\ &= \lambda\mu(\varphi\otimes\psi)(x,y). \end{aligned}$$

That is, the product $\lambda\mu$ is an eigenvalue of the eigenfunction $\varphi \otimes \psi$. ■

Solutions to selected Chapter 13 exercises.

13.3. For a directed edge $e = (z, w)$, we define $\nabla f(e) := f(w) - f(z)$. Observe that

$$2\tilde{\mathcal{E}}(f) = \sum_{(x,y)\in\tilde{E}} \tilde{Q}(x,y)[f(x) - f(y)]^2 = \sum_{x,y} \tilde{Q}(x,y) \sum_{\Gamma\in\mathcal{P}_{xy}} \nu_{xy}(\Gamma) \left[\sum_{e\in\Gamma} \nabla f(e)\right]^2.$$

Applying the Cauchy-Schwarz inequality yields

$$\begin{aligned} 2\tilde{\mathcal{E}}(f) &\leq \sum_{x,y} \tilde{Q}(x,y) \sum_{\Gamma\in\mathcal{P}_{xy}} \nu_{xy}(\Gamma)|\Gamma| \sum_{e\in\Gamma} [\nabla f(e)]^2 \\ &= \sum_{e\in E} [\nabla f(e)]^2 \sum_{(x,y)\in\tilde{E}} \tilde{Q}(x,y) \sum_{\Gamma: e\in\Gamma\in\mathcal{P}_{xy}} \nu_{xy}(\Gamma)|\Gamma|. \end{aligned}$$

By the definition of the congestion ratio, the right-hand side is bounded above by

$$\sum_{(z,w)\in E} BQ(z,w)[f(w) - f(z)]^2 = 2B\mathcal{E}(f),$$

completing the proof of (13.21).

The inequality (13.24) follows from Lemma 13.22. ∎

13.4. We compute the congestion ratio

$$B := \max_{e\in E} \left(\frac{1}{Q(e)} \sum_{(x,y)\in\tilde{E}} \tilde{Q}(x,y) \sum_{\Gamma: e\in\Gamma\in\mathcal{P}_{xy}} \nu_{xy}(\Gamma)|\Gamma| \right)$$

necessary to apply Corollary 13.26, following the outline of the proof of Corollary 13.27. To get a measure on paths between b and c, we write $c = ab$ and give weight $\nu_a(s_1, \ldots, s_k)$ to the path Γ_{bc} corresponding to $c = s_1 \cdots s_k b$.

For how many pairs $\{g, h\} \in \tilde{E}$ does a specific $e \in E$ appear in some Γ_{gh}, and with what weight does it appear? Let $s \in S$ be the generator corresponding to e, that is, $e = \{b, sb\}$ for some $b \in G$. For every occurrence of an edge $\{c, sc\}$ using s in some $\Gamma \in \mathcal{P}_a$, where $a \in \tilde{S}$, the edge e appears in the path $\Gamma_{c^{-1}b, ac^{-1}b} \in \mathcal{P}_{c^{-1}b, ac^{-1}b}$. Furthermore, $\nu_{c^{-1}b, ac^{-1}b}(\Gamma_{c^{-1}b, ac^{-1}b}) = \nu_a(\Gamma)$.

Hence the congestion ratio simplifies to

$$B = \max_{s\in S} \frac{1}{\mu(s)} \sum_{a\in\tilde{S}} \tilde{\mu}(a) \sum_{\Gamma\in\mathcal{P}_a} \nu_a(\Gamma) N(s,\Gamma)\, |\Gamma|.$$

∎

13.5. We bound $\binom{n}{\delta k} \leq n^{\delta k}/(\delta k)!$ and similarly bound $\binom{(1+\delta)k}{\delta k}$. Also, $\binom{n}{k} \geq n^k/k^k$. This gives

$$\sum_{k=1}^{n/2} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} \leq \sum_{k=1}^{n/2} \frac{n^{\delta k}((1+\delta)k)^{2\delta k} k^k}{(\delta k)!^3 n^k}.$$

Recall that for any integer ℓ we have $\ell! > (\ell/e)^\ell$, and we bound $(\delta k)!$ by this. We get

$$\sum_{k=1}^{n/2} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} \leq \sum_{k=1}^{\log n} \left(\frac{\log n}{n}\right)^{(1-\delta)k} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta k} + \sum_{k=\log n}^{n/2} \left(\frac{k}{n}\right)^{(1-\delta)k} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta k}.$$

The first sum clearly tends to 0 as n tends to ∞ for any $\delta \in (0,1)$. Since $k/n \leq 1/2$ and

$$(1/2)^{(1-\delta)} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta} < 0.8$$

for $\delta < 0.03$, for any such δ the second sum tends to 0 as n tends to ∞. ■

Solutions to selected Chapter 14 exercises.

14.2. If $\mathrm{Lip}(f) \leq 1$ and (X, Y) is a coupling of μ and ν attaining the minimum in the definition of transportation distance, then

$$\left|\int f d\mu - \int f d\nu\right| = |\mathbf{E}(f(X) - f(Y))| \leq \mathbf{E}(\rho(X,Y)) = \rho_K(\mu,\nu),$$

where we used $\mathrm{Lip}(f) \leq 1$ for the inequality and the fact that (X, Y) is the optimal coupling for the last equality. ■

14.3. We proceed by induction. Let H_j be the function defined in the first j steps described above; the domain of H_j is $[j]$. Clearly H_1 is uniform on $\Omega_{k,1}$. Suppose H_{j-1} is uniform on $\Omega_{k,j-1}$. Let $h \in \Omega_{k,j}$. Write h_{j-1} for the restriction of h to the domain $[j-1]$. Then

$$\mathbf{P}\{H_{j-1} = h_{j-1}\} = |\Omega_{k,j-1}|^{-1},$$

by the induction hypothesis. Note that

$$|\Omega_{k,j}| = (k-1)|\Omega_{k,j-1}|,$$

since for each element of $\Omega_{k,j-1}$ there are $k-1$ ways to extend it to an element of $\Omega_{k,j}$, and every element of $\Omega_{k,j}$ can be obtained as such an extension. By the construction and the induction hypothesis,

$$\begin{aligned}\mathbf{P}\{H_j = h\} &= \mathbf{P}\{H_{j-1} = h_{j-1}\}\mathbf{P}\{H_j = h \mid H_{j-1} = h_{j-1}\} \\ &= \frac{1}{|\Omega_{k,j-1}|}\frac{1}{(k-1)} \\ &= |\Omega_{k,j}|^{-1}.\end{aligned}$$

■

14.4. This is established by induction. The cases $n = 0$ and $n = 1$ are clear. Suppose it holds for $n \leq k-1$. The number of configurations $\omega \in \Omega_k$ with $\omega(k) = 0$ is the same as the total number of configurations in Ω_{k-1}. Also, the number of configurations $\omega \in \Omega_k$ with $\omega(k) = 1$ is the same as the number of configurations in Ω_{k-1} having no particle at $k-1$, which is the same as the number of configurations in Ω_{k-2}. ■

14.5. Let ω be an element of Ω_n, and let X be the random element of Ω_n generated by the algorithm. If $\omega(n) = 1$, then

$$\mathbf{P}\{X = \omega\} = \frac{1}{f_{n-1}}\left(\frac{f_{n-1}}{f_{n+1}}\right) = \frac{1}{f_{n+1}}.$$

Similarly, if $\omega(n) = 0$, then $\mathbf{P}\{X = \omega\} = 1/f_{n+1}$. ■

Solutions to selected Chapter 17 exercises.

17.1. Let (X_t) be simple random walk on $\mathbb{Z}$.

$$\begin{aligned} M_{t+1} - M_t &= (X_t + \Delta X_t)^3 - 3(t+1)(X_t + \Delta X_t) - X_t^3 + 3tX_t \\ &= 3X_t^2(\Delta X_t) + 3X_t(\Delta X_t)^2 + (\Delta X_t)^3 - 3t(\Delta X_t) - 3X_t - \Delta X_t. \end{aligned}$$

Note that $(\Delta X_t)^2 = 1$, so

$$M_{t+1} - M_t = (\Delta X_t)(3X_t^2 - 3t),$$

and

$$\mathbf{E}_k\left(M_{t+1} - M_t \mid X_t\right) = (3X_t^2 - 3t)\mathbf{E}_k(\Delta X_t \mid X_t) = 0.$$

Using the Optional Stopping Theorem,

$$\begin{aligned} k^3 &= \mathbf{E}_k(M_\tau) \\ &= \mathbf{E}_k\left[\left(X_\tau^3 - 3\tau X_\tau\right)\mathbf{1}_{\{X_\tau = n\}}\right] \\ &= n^3\mathbf{P}_k\{X_\tau = n\} - 3n\mathbf{E}_k\left(\tau\mathbf{1}_{\{X_\tau = n\}}\right). \end{aligned}$$

Dividing through by $kn^{-1} = \mathbf{P}_k\{X_\tau = n\}$ shows that

$$nk^2 = n^3 - 3n\mathbf{E}_k\left(\tau \mid X_\tau = n\right).$$

Rearranging,

$$\mathbf{E}_k\left(\tau \mid X_\tau = n\right) = \frac{n^2 - k^2}{3}.$$

The careful reader will notice that we have used the Optional Stopping Theorem without verifying its hypotheses! The application can be justified by applying it to $\tau \wedge B$ and then letting $B \to \infty$ and appealing to the Dominated Convergence Theorem. ■

17.2. Suppose that (X_t) is a supermartingale with respect to the sequence (Y_t). Define

$$A_t = -\sum_{s=1}^{t}\mathbf{E}(X_s - X_{s-1} \mid Y_0, \ldots, Y_{s-1}).$$

Since A_t is a function of $Y_0, \ldots, Y_{t-1}$, it is previsible. The supermartingale property ensures that

$$A_t - A_{t-1} = -\mathbf{E}(X_t - X_{t-1} \mid Y_0, \ldots, Y_{t-1}) \geq 0,$$

whence the sequence A_t is non-decreasing. Define $M_t := X_t + A_t$. Then

$$\begin{aligned} \mathbf{E}(M_{t+1} - M_t \mid Y_0, \ldots, Y_t) &= \mathbf{E}(X_{t+1} - X_t \mid Y_0, \ldots, Y_t) \\ &\quad - \mathbf{E}(\mathbf{E}(X_{t+1} - X_t \mid Y_0, \ldots, Y_t) \mid Y_0, \ldots, Y_t) \\ &= 0. \end{aligned}$$

■

17.3. Using the Doob decomposition, $Z_t = M_t - A_t$, where (M_t) is a martingale with $M_0 = Z_0$ and (A_t) is a previsible and non-decreasing sequence with $A_0 = 0$.

Note that since both Z_t and A_t are non-negative, so is (M_t). Furthermore,

$$A_{t+1} - A_t = \mathbf{E}(Z_{t+1} - Z_t \mid \mathbf{Y}_t) \leq B,$$

so

$$\mathbf{E}(M_{t+1} - M_t \mid \mathbf{Y}_t) \leq \mathbf{E}(Z_{t+1} - Z_t \mid \mathbf{Y}_t) + B \leq 2B.$$

Since (A_t) is previsible, on the event that $\tau > t$,

$$\mathrm{Var}(M_{t+1} \mid Y_1, \ldots, Y_t) = \mathrm{Var}(Z_{t+1} \mid Y_1, \ldots, Y_t) \geq \sigma^2 > 0. \tag{C.24}$$

Given $h \geq 2B$, consider the stopping time

$$\tau_h = \min\{t : M_t \geq h\} \wedge \tau \wedge u.$$

Since τ_h is bounded by u, the Optional Stopping Theorem yields

$$k = \mathbf{E}(M_{\tau_h}) \geq h\mathbf{P}\{M_{\tau_h} \geq h\}.$$

Rearranging, we have that

$$\mathbf{P}\{M_{\tau_h} \geq h\} \leq \frac{k}{h}. \tag{C.25}$$

Let

$$W_t := M_t^2 - hM_t - \sigma^2 t.$$

The inequality (C.24) implies that $\mathbf{E}(W_{t+1} \mid \mathbf{Y}_t) \geq W_t$ whenever $\tau > t$. That is, $W_{t\wedge\tau}$ is a submartingale. By optional stopping, since τ_h is bounded and $\tau_h \wedge \tau = \tau_h$,

$$-kh \leq \mathbf{E}(W_0) \leq \mathbf{E}(W_{\tau_h}) = \mathbf{E}\left(M_{\tau_h}(M_{\tau_h} - h)\right) - \sigma^2\mathbf{E}(\tau_h).$$

Since $M_{\tau_h}(M_{\tau_h} - h)$ is non-positive on the event $M_{\tau_h} \leq h$, the right-hand side above is bounded above by

$$(h + 2B)(2B)\mathbf{P}\{\tau_h > h\} + \sigma^2\mathbf{E}(\tau_h) \leq 2h^2\mathbf{P}\{\tau_h > h\} - \sigma^2\mathbf{E}(\tau_h).$$

Combining these two bounds and using (C.25) shows that $\sigma^2\mathbf{E}(\tau_h) \leq kh + 2h^2(k/h) = 3kh$. Therefore,

$$\begin{aligned}\mathbf{P}\{\tau > u\} &\leq \mathbf{P}\{M_{\tau_h} \geq h\} + \mathbf{P}\{\tau_h \geq u\}\\ &\leq \frac{k}{h} + \frac{3kh}{u\sigma^2},\end{aligned}$$

using Markov's inequality and the bound on $\mathbf{E}(\tau_h)$ in the last step.

Optimize by choosing $h = \sqrt{u\sigma^2/3}$, obtaining

$$\mathbf{P}\{\tau > u\} \leq \frac{2\sqrt{3}k}{\sigma\sqrt{u}} \leq \frac{4k}{\sigma\sqrt{u}}. \tag{C.26}$$

■

Solution to Chapter 18 exercise.

18.1. First suppose that the chain satisfies (18.24). Then for any $\gamma > 0$, for n large enough,

$$t_{\text{mix}}(\varepsilon) \leq (1+\gamma)t^n_{\text{mix}},$$
$$t_{\text{mix}}(1-\varepsilon) \geq (1-\gamma)t^n_{\text{mix}}.$$

Thus

$$\frac{t_{\text{mix}}(\varepsilon)}{t_{\text{mix}}(1-\varepsilon)} \leq \frac{1+\gamma}{1-\gamma}.$$

Letting $\gamma \downarrow 0$ shows that (18.3) holds.

Suppose that (18.3) holds. Fix $\gamma > 0$. For any $\varepsilon > 0$, for n large enough, $t_{\text{mix}}(\varepsilon) \leq (1+\gamma)t^n_{\text{mix}}$. That is, $\lim_{n\to\infty} d_n\left((1+\gamma)t^n_{\text{mix}}\right) \leq \varepsilon$. Since this holds for all ε,

$$\lim_{n\to\infty} d_n\left((1+\gamma)t^n_{\text{mix}}\right) = 0.$$

Also, $\lim_{n\to\infty} d_n\left((1-\gamma)t^n_{\text{mix}}\right) \geq 1-\varepsilon$, since $t_{\text{mix}}(1-\varepsilon) \geq (1-\gamma)t^n_{\text{mix}}$ for n sufficiently large. Consequently,

$$\lim_{n\to\infty} d_n\left((1-\gamma)t^n_{\text{mix}}\right) = 1.$$

■

Solutions to selected Chapter 20 exercises.

20.1. The distribution of a sum of n independent exponential random variables with rate μ has a Gamma distribution with parameters n and μ, so S_k has density

$$f_k(s) = \frac{\mu^k s^{k-1} e^{-\mu s}}{(k-1)!}.$$

Since S_k and X_{k+1} are independent,

$$\begin{aligned}\mathbf{P}\{S_k \leq t < S_k + X_{k+1}\} &= \int_0^t \frac{\mu^k s^{k-1} e^{-\mu s}}{(k-1)!} \int_{t-s}^{\infty} \mu e^{-\mu x}\,dx\,ds \\ &= \int_0^t \frac{\mu^k s^{k-1}}{(k-1)!} e^{-\mu t}\,ds \\ &= \frac{(\mu t)^k e^{-\mu t}}{k!}.\end{aligned}$$

■

20.3. From the definition of e^{A+B},

$$e^{A+B} = \sum_{n=0}^{\infty} \frac{(A+B)^n}{n!}. \tag{C.27}$$

Since A and B commute, $(A+B)^n$ has a binomial formula:

$$(A+B)^n = \sum_{k=0}^{n} \binom{n}{k} A^n B^{n-k}.$$

Therefore, the left-hand side of (C.27) equals

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{A^k}{k!}\frac{B^{n-k}}{(n-k)!} = \sum_{k=0}^{\infty}\frac{A^k}{k!}\sum_{j=0}^{\infty}\frac{A^j}{j!} = e^A e^B.$$

■

20.5. Let $\Omega = \prod_{i=1}^{n} \Omega_i$. We have

$$I(\mu, \nu) = \sum_{\boldsymbol{x} \in \Omega} \sqrt{\mu(\boldsymbol{x})\nu(\boldsymbol{y})} = \sum_{x_1 \in \Omega_1} \cdots \sum_{x_n \in \Omega_n} \sqrt{\prod_{i=1}^{n} \mu_i(x_i) \prod_{i=1}^{n} \nu_i(x_i)}$$
$$= \left[\sum_{x_1 \in \Omega_1} \sqrt{\mu_1(x_1)\nu_1(x_1)} \right] \cdots \left[\sum_{x_n \in \Omega_n} \sqrt{\mu_n(x_n)\nu_n(x_n)} \right] = \prod_{i=1}^{n} I(\mu_i, \nu_i).$$

■

Solutions to selected Chapter 21 exercises.

21.1. We can write $X_t = x + \sum_{s=1}^{t} Y_s$, where $x \in \Omega$ and $(Y_s)_{s=1}^{\infty}$ is an i.i.d. sequence of $\{-1, 1\}$-valued random variables satisfying

$$\mathbf{P}\{Y_s = +1\} = p,$$
$$\mathbf{P}\{Y_s = -1\} = q.$$

By the Strong Law, $\mathbf{P}_0\{\lim_{t\to\infty} t^{-1} X_t = (p - q)\} = 1$. In particular,

$$\mathbf{P}_0\{X_t > (p - q)t/2 \text{ for } t \text{ sufficiently large}\} = 1.$$

That is, with probability one, there are only finitely many visits of the walker to 0. Since the number of visits to 0 is a geometric random variable with parameter $\mathbf{P}_0\{\tau_0^+ = \infty\}$ (see the proof of Proposition 21.3), this probability must be positive. ■

21.2. Suppose that $\pi(v) = 0$. Since $\pi = \pi P$,

$$0 = \pi(v) = \sum_{u \in X} \pi(u) P(u, v).$$

Since all the terms on the right-hand side are non-negative, each is zero. That is, if $P(u, v) > 0$, it must be that $\pi(u) = 0$.

Suppose that there is some $y \in \Omega$ so that $\pi(y) = 0$. By irreducibility, for any $x \in \Omega$, there is a sequence $u_0, \ldots, u_t$ so that $u_0 = x$, $u_t = y$, and each $P(u_{i-1}, u_i) > 0$ for $i = 1, \ldots, t$. Then by induction it is easy to see that $\pi(u_i) = 0$ for each of $i = 0, 1, 2, \ldots, t$. Thus $\pi(x) = 0$ for all $x \in \Omega$, and π is not a probability distribution. ■

21.4. If the original graph is regarded as a network with conductances $c(e) = 1$ for all e, then the subgraph is also a network, but with $c(e) = 0$ for all edges which are omitted. By Rayleigh's Monotonicity Law, the effective resistance from a fixed vertex v to ∞ is not smaller in the subgraph than for the original graph. This together with Proposition 21.6 shows that the subgraph must be recurrent. ■

21.5. Define

$$A_{x,y} = \{t : P^t(x, y) > 0\}.$$

By aperiodicity, g.c.d.$(A_{x,x}) = 1$. Since $A_{x,x}$ is closed under addition, there is some t_x so that $t \in A_{x,x}$ for $t \geq t_x$. Also, by irreducibility, there is some s so that $P^s(x, y) > 0$. Since

$$P^{t+s}(x, y) \geq P^t(x, x) P^s(x, y),$$

if $t \geq t_x$, then $t+s \in A_{y,x}$. That is, there exists $t_{x,y}$ such that if $t \geq t_{x,y}$, then $t \in A_{x,y}$.

Let $t_0 = \max\{t_{x,z}, t_{y,w}\}$. If $t \geq t_0$, then $P^t(x,z) > 0$ and $P^t(y,w) > 0$. In particular,

$$P^{t_0}((x,y),(z,w)) = P^{t_0}(x,z)P^{t_0}(y,w) > 0.$$

■

21.6. (X_t) is a nearest-neighbor random walk on $\mathbb{Z}^+$ which increases by 1 with probability α and decreases by 1 with probability $\beta = 1-\alpha$. When the walker is at 0, instead of decreasing with probability β, it remains at 0. Thus if $\alpha < \beta$, then the chain is a downwardly biased random walk on $\mathbb{Z}^+$, which was shown in Example 21.15 to be positive recurrent.

If $\alpha = \beta$, this is an unbiased random walk on $\mathbb{Z}^+$. This is null recurrent for the same reason that the simple random walk on $\mathbb{Z}$ is null recurrent, shown in Example 21.10.

Consider the network with $V = \mathbb{Z}^+$ and with $c(k,k+1) = r^k$. If $r = p/(1-p)$, then the random walk on the network corresponds to a nearest-neighbor random walk which moves "up" with probability p. The effective resistance from 0 to n is

$$\mathcal{R}(0 \leftrightarrow n) = \sum_{k=1}^{n} r^{-k}.$$

If $p > 1/2$, then $r > 1$ and the right-hand side converges to a finite number, so $\mathcal{R}(0 \leftrightarrow \infty) < \infty$. By Proposition 21.6 this walk is transient. The FIFO queue of this problem is an upwardly biased random walk when $\alpha > \beta$, and thus it is transient as well. ■

21.7. Let $r = \alpha/\beta$. Then $\pi(k) = (1-r)r^k$ for all $k \geq 0$, that is, π is the geometric distribution with probability r shifted by 1 to the left. Thus

$$E_\pi(X+1) = 1/(1-r) = \beta/(\beta-\alpha).$$

Since $\mathbf{E}(T \mid X \text{ before arrival}) = (1+X)/\beta$, we conclude that $\mathbf{E}_\pi(T) = 1/(\beta - \alpha)$. ■

21.8. Suppose that $\mu = \mu P$, so that for all k,

$$\mu(k) = \frac{\mu(k-1) + \mu(k+1)}{2}.$$

The difference sequence $d(k) = \mu(k) - \mu(k-1)$ is easily seen to be constant, and hence μ is not bounded. ■

Solutions to selected Appendix B exercises.

B.4. Let $g(y,u)$ be the joint density of (Y, U_Y). Then

$$\begin{aligned} f_{Y,U}(y,u) &= f_Y(y) f_{U_Y|Y}(u|\, y) \\ &= g(y)\mathbf{1}\{g(y)>0\}\frac{\mathbf{1}\{0 \leq u \leq Cg(y)\}}{Cg(y)} = \frac{1}{C}\mathbf{1}\{g(y)>0, u \leq Cg(y)\}. \end{aligned} \tag{C.28}$$

This is the density for a point (Y,U) drawn from the region under the graph of the function g.

Conversely, let (Y, U) be a uniform point from the region under the graph of the function g. Its density is the right-hand side of (C.28). The marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} \frac{1}{C}\mathbf{1}\{g(y) > 0, u \le Cg(y)\}du = \mathbf{1}\{g(y) > 0\}\frac{1}{C}Cg(y) = g(y). \quad \text{(C.29)}$$

■

B.9. Let R be any region of TA. First, note that since $\text{rank}(T) = d$, by the Rank Theorem, T is one-to-one. Consequently, $TT^{-1}R = R$, and

$$\text{Volume}_d(R) = \text{Volume}_d(TT^{-1}R) = \sqrt{\det(T^tT)}\,\text{Volume}(T^{-1}R),$$

so that $\text{Volume}(T^{-1}R) = \text{Volume}_d(R)/\sqrt{\det(T^tT)}$. To find the distribution of Y, we compute

$$\mathbf{P}\{Y \in R\} = \mathbf{P}\{TX \in R\} = \mathbf{P}\{X \in T^{-1}R\}. \quad \text{(C.30)}$$

Since X is uniform, the right-hand side is

$$\frac{\text{Volume}(T^{-1}R)}{\text{Volume}(A)} = \frac{\text{Volume}_d(R)}{\sqrt{\det(T^tT)}\,\text{Volume}(A)} = \frac{\text{Volume}_d(R)}{\text{Volume}_d(TA)}. \quad \text{(C.31)}$$

■

B.11.

(a) $x \le U_{(k)} \le x + dx$ if and only if among $\{U_1, U_2, \ldots, U_n\}$ exactly $k-1$ lie to the left of x, one is in $[x, x+dx]$, and $n-k$ variables exceed $x + dx$. This occurs with probability

$$\binom{n}{(k-1), 1, (n-k)} x^{k-1}(1-x)^{n-k}dx.$$

Thus,

$$\begin{aligned}
\mathbf{E}\left(U_{(k)}\right) &= \int_0^1 \frac{n!}{(k-1)!(n-k)!} x^k (1-x)^{n-k} dx \\
&= \frac{n!}{(k-1)!(n-k)!}\frac{(n-k)!k!}{(n+1)!} \\
&= \frac{k}{n+1}.
\end{aligned}$$

(The integral can be evaluated by observing that the function

$$x \mapsto \frac{k!(n-k)!}{(n+1)!}x^k(1-x)^{n-k}$$

is the density for a Beta random variable with parameters $k+1$ and $n-k+1$.)

(b) The distribution function for $U_{(n)}$ is

$$F_n(x) = \mathbf{P}\{U_1 \le x, U_2 \le x, \ldots, U_n \le x\} = \mathbf{P}\{U_1 \le x\}^n = x^n.$$

Differentiating, the density function for $U_{(n)}$ is

$$f_n(x) = nx^{n-1}.$$

Consequently,

$$\mathbf{E}\left(U_{(n)}\right) = \int_0^1 xnx^{n-1}dx = \frac{n}{n+1}x^{n+1}\Big|_0^1 = \frac{n}{n+1}.$$

We proceed by induction, showing that

$$\mathbf{E}\left(U_{(n-k)}\right) = \frac{n-k}{n+1}. \tag{C.32}$$

We just established the case $k = 0$. Now suppose (C.32) holds for $k = j$. Given $U_{(n-j)}$, the order statistics $U_{(i)}$ for $i = 1, \dots, n-j-1$ have the distribution of the order statistics for $n-j-1$ independent variables uniform on $[0, U_{(n-j)}]$. Thus,

$$\mathbf{E}\left(U_{(n-j-1)} \mid U_{(n-j)}\right) = U_{(n-j)} \frac{n-j-1}{n-j},$$

and so

$$\mathbf{E}\left(U_{(n-j-1)}\right) = \mathbf{E}\left(\mathbf{E}\left(U_{(n-j-1)} \mid U_{(n-j)}\right)\right) = \mathbf{E}\left(U_{(n-j)}\right) \frac{n-j-1}{n-j}.$$

Since (C.32) holds for $k = j$ by assumption,

$$\mathbf{E}\left(U_{(n-j-1)}\right) = \frac{n-j}{n+1}\frac{n-j-1}{n-j} = \frac{n-j-1}{n+1}.$$

This establishes (C.32) for $j = k$.

(c) The joint density of $(S_1, S_2, \dots, S_{n+1})$ is $e^{-s_{n+1}}\mathbf{1}_{\{0<s_1<\cdots<s_{n+1}\}}$, as can be verified by induction:

$$\begin{aligned} f_{S_1,S_2,\dots,S_{n+1}}(s_1,\dots,s_{n+1}) &= f_{S_1,S_2,\dots,S_n}(s_1,\dots,s_n) f_{S_{n+1}|S_1,\dots,S_n}(s_{n+1} \mid s_1,\dots,s_n) \\ &= e^{-s_n}\mathbf{1}_{\{0<s_1<\cdots<s_n\}} e^{-(s_{n+1}-s_n)}\mathbf{1}_{\{s_n<s_{n+1}\}} \\ &= e^{-s_{n+1}}\mathbf{1}_{\{0<s_1<\cdots<s_{n+1}\}}. \end{aligned}$$

Because the density of S_{n+1} is $s_{n+1}^n e^{-s_{n+1}}/(n!)\mathbf{1}_{\{s_{n+1}>0\}}$,

$$f_{S_1,\dots,S_n|S_{n+1}}(s_1,\dots,s_n \mid s_{n+1}) = \frac{n!}{s_{n+1}^n}\mathbf{1}_{\{0<s_1<\cdots<s_n<s_{n+1}\}}.$$

If $T_k = S_k/S_{n+1}$ for $k = 1, \dots, n$, then

$$f_{T_1,\dots,T_k|S_{n+1}}(t_1,\dots,t_n \mid s_{n+1}) = n!\mathbf{1}_{\{0<t_1<\cdots<t_n<1\}}.$$

Since the right-hand side does not depend on s_{n+1}, the vector

$$\left(\frac{S_1}{S_{n+1}}, \frac{S_2}{S_{n+1}}, \dots, \frac{S_1}{S_{n+1}}\right)$$

is uniform over the set

$$\{(x_1,\dots,x_n) \,:\, x_1 < x_2 < \cdots < x_n\}.$$

■

B.14. We proceed by induction on n. The base case $n = 1$ is clear. Assume that the $(n-1)$-step algorithm indeed produces a uniformly distributed $\xi_{n-1} \in \Xi_{n-1}^{\mathrm{nr}}$. Extend ξ_{n-1} to ξ_n according to the algorithm, picking one of the three available extensions at random. Note that $|\Xi_n^{\mathrm{nr}}| = 4 \cdot 3^{n-1}$. For h any path in Ξ_n^{nr}, let h_{n-1} be the projection of h to Ξ_{n-1}^{nr}, and observe that

$$\begin{aligned} \mathbf{P}\{\xi_n = h\} &= \mathbf{P}\{\xi_n = h \mid \xi_{n-1} = h_{n-1}\}\mathbf{P}\{\xi_{n-1} = h_{n-1}\} \\ &= \frac{1}{3}\left(\frac{1}{4 \cdot 3^{n-2}}\right) = \frac{1}{4 \cdot 3^{n-1}}. \end{aligned}$$

■

B.15. Since the number of self-avoiding walks of length n is clearly bounded by $c_{n,4}$ and our method for generating non-reversing paths is uniform over Ξ_n^{nr} which has size $4 \cdot 3^{n-1}$, the second part follows from the first.

There are $4(3^3) - 8$ walks of length 4 starting at the origin which are non-reversing and do not return to the origin. At each 4-step stage later in the walk, there are 3^4 non-reversing paths of length 4, of which six create loops. This establishes (B.26). ■

Bibliography

The pages on which a reference appears follow the symbol ↑.

Ahlfors, L. V. 1978. *Complex analysis*, 3rd ed., McGraw-Hill Book Co., New York. An introduction to the theory of analytic functions of one complex variable; International Series in Pure and Applied Mathematics. ↑125

Aldous, D. J. 1983a. *On the time taken by random walks on finite groups to visit every state*, Z. Wahrsch. Verw. Gebiete **62**, no. 3, 361–374. ↑152

Aldous, D. 1983b. *Random walks on finite groups and rapidly mixing Markov chains*, Seminar on probability, XVII, Lecture Notes in Math., vol. 986, Springer, Berlin, pp. 243–297. ↑60, 218

Aldous, D. 1989a. *An introduction to covering problems for random walks on graphs*, J. Theoret. Probab. **2**, no. 1, 87–89. ↑152

Aldous, D. J. 1989b. *Lower bounds for covering times for reversible Markov chains and random walks on graphs*, J. Theoret. Probab. **2**, no. 1, 91–100. ↑152

Aldous, D. 1990. *A random walk construction of uniform spanning trees and uniform labelled trees*, SIAM Journal on Discrete Mathematics **3**, 450–465. ↑297

Aldous, D. J. 1991a. *Random walk covering of some special trees*, J. Math. Anal. Appl. **157**, no. 1, 271–283. ↑152

Aldous, D. J. 1991b. *Threshold limits for cover times*, J. Theoret. Probab. **4**, no. 1, 197–211. ↑152, 264

Aldous, D. 1995. *On simulating a Markov chain stationary distribution when transition probabilities are unknown* (D. Aldous, P. Diaconis, J. Spencer, and J. M. Steele, eds.), IMA Volumes in Mathematics and its Applications, vol. 72, Springer-Verlag. ↑296

Aldous, D. J. 1999. unpublished note. ↑180

Aldous, D. J. 2004, American Institute of Mathematics (AIM) research workshop "Sharp Thresholds for Mixing Times" (Palo Alto, December 2004). Summary available at `http://www.aimath.org/WWN/mixingtimes`. ↑255

Aldous, D. and P. Diaconis. 1986. *Shuffling cards and stopping times*, Amer. Math. Monthly **93**, no. 5, 333–348. ↑60, 85, 96, 112, 113

Aldous, D. and P. Diaconis. 1987. *Strong uniform times and finite random walks*, Adv. in Appl. Math. **8**, no. 1, 69–97. ↑85, 260

Aldous, D. and P. Diaconis. 2002. *The asymmetric one-dimensional constrained Ising model: rigorous results*, J. Statist. Phys. **107**, no. 5-6, 945–975. ↑98

Aldous, D. and J. Fill. 1999. *Reversible Markov chains and random walks on graphs*, in progress. Manuscript available at `http://www.stat.berkeley.edu/~aldous/RWG/book.html`. ↑xvi, 20, 60, 85, 136, 141, 144, 263, 299

Alon, N. 1986. *Eigenvalues and expanders*, Combinatorica **6**, no. 2, 83–96. ↑188

Alon, N. and V. D. Milman. 1985. λ_1, *isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B **38**, no. 1, 73–88. ↑188

Anantharam, V. and P. Tsoucas. 1989. *A proof of the Markov chain tree theorem*, Statistics and Probability Letters **8**, 189–192. ↑

Angel, O., Y. Peres, and D. B. Wilson. 2008. *Card shuffling and Diophantine approximation*, Ann. Appl. Probab. **18**, no. 3, 1215–1231. ↑168

Archer, A. F. 1999. *A modern treatment of the 15 puzzle*, Amer. Math. Monthly **106**, no. 9, 793–799. ↑336

Artin, M. 1991. *Algebra*, Prentice Hall Inc., Englewood Cliffs, NJ. ↑35, 111

Asmussen, S., P. Glynn, and H. Thorisson. 1992. *Stationary detection in the initial transient problem*, ACM Transactions on Modeling and Computer Simulation **2**, 130–157. ↑296

Barlow, M. T., T. Coulhon, and T. Kumagai. 2005. *Characterization of sub-Gaussian heat kernel estimates on strongly recurrent graphs*, Comm. Pure Appl. Math. **58**, no. 12, 1642–1677. ↑284

Barrera, J., B. Lachaud, and B. Ycart. 2006. *Cut-off for n-tuples of exponentially converging processes*, Stochastic Process. Appl. **116**, no. 10, 1433–1446. ↑274

Basharin, G. P., A. N. Langville, and V. A. Naumov. 2004. *The life and work of A. A. Markov*, Linear Algebra Appl. **386**, 3–26. ↑20

Baxter, R. J. 1982. *Exactly Solved Models in Statistical Mechanics*, Academic Press. ↑288

Bayer, D. and P. Diaconis. 1992. *Trailing the dovetail shuffle to its lair*, Ann. Appl. Probab. **2**, no. 2, 294–313. ↑111, 113

Benjamin, A. T. and J. J. Quinn. 2003. *Proofs that really count: The art of combinatorial proof*, Dolciani Mathematical Expositions, vol. 27, Math. Assoc. Amer., Washington, D. C. ↑196

Berger, N., C. Kenyon, E. Mossel, and Y. Peres. 2005. *Glauber dynamics on trees and hyperbolic graphs*, Probab. Theory Related Fields **131**, no. 3, 311–340. ↑214

Billingsley, P. 1995. *Probability and measure*, 3rd ed., Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York. ↑303, 308

Borovkov, A. A. and S. G. Foss. 1992. *Stochastically recursive sequences and their generalizations*, Siberian Advances in Mathematics **2**, 16–81. ↑288

Bodineau, T. 2005. *Slab percolation for the Ising model*, Probab. Theory Related Fields **132**, no. 1, 83–118. ↑215

Brémaud, P. 1999. *Markov chains*, Texts in Applied Mathematics, vol. 31, Springer-Verlag, New York. Gibbs fields, Monte Carlo simulation, and queues. ↑45

Broder, A. 1989. *Generating random spanning trees*, 30th Annual Symposium on Foundations of Computer Science, pp. 442–447. ↑297

Broder, A. Z. and A. R. Karlin. 1989. *Bounds on the cover time*, J. Theoret. Probab. **2**, no. 1, 101–120. ↑152

Bubley, R. and M. Dyer. 1997. *Path coupling: A technique for proving rapid mixing in Markov chains*, Proceedings of the 38th Annual Symposium on Foundations of Computer Science, pp. 223–231. ↑191

Cancrini, N., P. Caputo, and F. Martinelli. 2006. *Relaxation time of L-reversal chains and other chromosome shuffles*, Ann. Appl. Probab. **16**, no. 3, 1506–1527. ↑227

Cancrini, N., F. Martinelli, C. Roberto, and C. Toninelli. 2008. *Kinetically constrained spin models*, Probab. Theory Related Fields **140**, no. 3-4, 459–504, available at `arXiv:math/0610106v1`. ↑98

Cerf, R. and A. Pisztora. 2000. *On the Wulff crystal in the Ising model*, Ann. Probab. **28**, no. 3, 947–1017. ↑215

Cesi, F., G. Guadagni, F. Martinelli, and R. H. Schonmann. 1996. *On the two-dimensional stochastic Ising model in the phase coexistence region near the critical point*, J. Statist. Phys. **85**, no. 1-2, 55–102. ↑215

Chandra, A. K., P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari. 1996/97. *The electrical resistance of a graph captures its commute and cover times*, Comput. Complexity **6**, no. 4, 312–340. Extended abstract originally published in *Proc. 21st ACM Symp. Theory of Computing* (1989) 574–586. ↑141

Chayes, J. T., L. Chayes, and R. H. Schonmann. 1987. *Exponential decay of connectivities in the two-dimensional Ising model*, J. Statist. Phys. **49**, no. 3-4, 433–445. ↑214

Cheeger, J. 1970. *A lower bound for the smallest eigenvalue of the Laplacian*, Problems in analysis (Papers dedicated to Salomon Bochner, 1969), Princeton Univ. Press, Princeton, pp. 195–199. ↑187

Chen, F., L. Lovász, and I. Pak. 1999. *Lifting Markov chains to speed up mixing*, Annual ACM Symposium on Theory of Computing (Atlanta, GA, 1999), ACM, New York, pp. 275–281 (electronic). ↑98

Chen, G.-Y. and L. Saloff-Coste. 2008. *The cutoff phenomenon for ergodic Markov processes*, Electron. J. Probab. **13**, no. 3, 26–78. ↑255

Chen, M.-F. 1998. *Trilogy of couplings and general formulas for lower bound of spectral gap*, Probability towards 2000 (New York, 1995), Lecture Notes in Statist., vol. 128, Springer, New York, pp. 123–136. ↑171, 258

Chung, F., P. Diaconis, and R. Graham. 2001. *Combinatorics for the East model*, Adv. in Appl. Math. **27**, no. 1, 192–206. ↑98

Chyakanavichyus, V. and P. Vaĭtkus. 2001. *Centered Poisson approximation by the Stein method*, Liet. Mat. Rink. **41**, no. 4, 409–423 (Russian, with Russian and Lithuanian summaries); English transl.,. 2001, Lithuanian Math. J. **41**, no. 4, 319–329. ↑274

Dembo, A., Y. Peres, J. Rosen, and O. Zeitouni. 2004. *Cover times for Brownian motion and random walk in two dimensions*, Ann. Math. **160**, 433–464. ↑152

Devroye, L. 1986. *Nonuniform random variate generation*, Springer-Verlag, New York. ↑325

Diaconis, P. 1988. *Group Representations in Probability and Statistics*, Lecture Notes - Monograph Series, vol. 11, Inst. Math. Stat., Hayward, CA. ↑35, 102, 104, 113, 156, 168

Diaconis, P. 1996. *The cutoff phenomenon in finite Markov chains*, Proc. Nat. Acad. Sci. U.S.A. **93**, no. 4, 1659–1664. ↑

Diaconis, P. 2003. *Mathematical developments from the analysis of riffle shuffling*, Groups, combinatorics & geometry (Durham, 2001), World Sci. Publ., River Edge, NJ, pp. 73–97. ↑113

Diaconis, P. and J. A. Fill. 1990. *Strong stationary times via a new form of duality*, Ann. Probab. **18**, no. 4, 1483–1522. ↑85, 168, 243, 255

Diaconis, P. and D. Freedman. 1999. *Iterated random functions*, SIAM Review **41**, 45–76. ↑288

Diaconis, P., M. McGrath, and J. Pitman. 1995. *Riffle shuffles, cycles, and descents*, Combinatorica **15**, no. 1, 11–29. ↑113

Diaconis, P. and L. Saloff-Coste. 1993a. *Comparison theorems for reversible Markov chains*, Ann. Appl. Probab. **3**, no. 3, 696–730. ↑182, 188

Diaconis, P. and L. Saloff-Coste. 1993b. *Comparison techniques for random walk on finite groups*, Ann. Probab. **21**, no. 4, 2131–2156. ↑188, 227

Diaconis, P. and L. Saloff-Coste. 1996. *Nash inequalities for finite Markov chains*, J. Theoret. Probab. **9**, no. 2, 459–510. ↑141

Diaconis, P. and L. Saloff-Coste. 1998. *What do we know about the Metropolis algorithm?*, J. Comput. System Sci. **57**, no. 1, 20–36. 27th Annual ACM Symposium on the Theory of Computing (STOC'95) (Las Vegas, NV). ↑45

Diaconis, P. and L. Saloff-Coste. 2006. *Separation cut-offs for birth and death chains*, Ann. Appl. Probab. **16**, no. 4, 2098–2122. ↑255, 256, 301

Diaconis, P. and M. Shahshahani. 1981. *Generating a random permutation with random transpositions*, Z. Wahrsch. Verw. Gebiete **57**, no. 2, 159–179. ↑102, 168, 227

Diaconis, P. and D. Stroock. 1991. *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab. **1**, no. 1, 36–61. ↑182, 188

Ding, J., E. Lubetzky, and Y. Peres. 2008a. *The mixing time evolution of Glauber dynamics for the mean-field Ising model*, available at `arXiv:0806.1906`. ↑214

Ding, J., E. Lubetzky, and Y Peres. 2008b. *Total-variation cutoff in birth-and-death chains*, available at `arXiv:0801.2625`. ↑255, 301

Dobrushin, R., R. Kotecký, and S. Shlosman. 1992. *Wulff construction. A global shape from local interaction*, Translations of Mathematical Monographs, vol. 104, American Mathematical Society, Providence, RI. Translated from the Russian by the authors. ↑215

Dobrushin, R. L. and S. B. Shlosman. 1987. *Completely analytical interactions: constructive description*, J. Statist. Phys. **46**, no. 5-6, 983–1014. ↑215

Doeblin, W. 1938. *Esposé de la théorie des chaînes simple constantes de Markov à un nombre fini d'états*, Rev. Math. Union Interbalkan. **2**, 77–105. ↑74

Doob, J. L. 1953. *Stochastic processes*, John Wiley & Sons Inc., New York. ↑245

Doyle, P. G. and E. J. Snell. 1984. *Random walks and electrical networks*, Carus Math. Monographs, vol. 22, Math. Assoc. Amer., Washington, D. C. ↑125, 285, 286

Dubins, L. E. 1968. *On a theorem of Skorohod*, Ann. Math. Statist. **39**, 2094–2097. ↑85

Dudley, R. M. 2002. *Real analysis and probability*, Cambridge Studies in Advanced Mathematics, vol. 74, Cambridge University Press, Cambridge. Revised reprint of the 1989 original. ↑200

Durrett, R. 2003. *Shuffling chromosomes*, J. Theoret. Probab. **16**, 725–750. ↑221, 227

Durrett, R. 2005. *Probability: theory and examples*, third edition, Brooks/Cole, Belmont, CA. ↑267, 303

Dyer, M., L. A. Goldberg, and M. Jerrum. 2006a. *Dobrushin conditions and systematic scan*, Approximation, randomization and combinatorial optimization, Lecture Notes in Comput. Sci., vol. 4110, Springer, Berlin, pp. 327–338. ↑300

Dyer, M., L. A. Goldberg, and M. Jerrum. 2006b. *Systematic scan for sampling colorings*, Ann. Appl. Probab. **16**, no. 1, 185–230. ↑300

Dyer, M., L. A. Goldberg, M. Jerrum, and R. Martin. 2006. *Markov chain comparison*, Probab. Surv. **3**, 89–111 (electronic). ↑

Dyer, M. and C. Greenhill. 2000. *On Markov chains for independent sets*, J. Algorithms **35**, no. 1, 17–49. ↑295

Dyer, M., C. Greenhill, and M. Molloy. 2002. *Very rapid mixing of the Glauber dynamics for proper colorings on bounded-degree graphs*, Random Structures Algorithms **20**, no. 1, 98–114. ↑199

Eggenberger, F. and G Pólya. 1923. *Über die Statistik vorketter vorgänge*, Zeit. Angew. Math. Mech. **3**, 279–289. ↑35

Elias, P. 1972. *The efficient construction of an unbiased random sequence*, Ann. Math. Statist. **43**, 865–870. ↑325

Einstein, A. 1934. *On the method of theoretical physics*, Philosophy of Science **1**, no. 2, 163–169. ↑1

Feller, W. 1968. *An introduction to probability theory and its applications*, third edition, Vol. 1, Wiley, New York. ↑20, 35, 168, 286, 303

Fill, J. A. 1991. *Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process*, Ann. Appl. Probab. **1**, no. 1, 62–87. ↑98

Fill, J. A. 1998. *An interruptible algorithm for perfect sampling via Markov chains*, Annals of Applied Probability **8**, 131–162. ↑288, 292

Fill, J. A. 2007. *On hitting times and fastest strong stationary times for skip-free chains*, available at `arXiv:0708.4258v1[math.PR]`. ↑168

Fill, J. A. and M. Huber. 2000. *The randomness recycler: A new technique for perfect sampling*, 41st Annual Symposium on Foundations of Computer Science, pp. 503–511. ↑288

Fill, J. A., M. Machida, D. J. Murdoch, and J. S. Rosenthal. 2000. *Extension of Fill's perfect rejection sampling algorithm to general chains*, Random Structure and Algorithms **17**, 290–316. ↑288

Frieze, A. and E. Vigoda. 2007. *A survey on the use of Markov chains to randomly sample colourings*, Combinatorics, complexity, and chance, Oxford Lecture Ser. Math. Appl., vol. 34, Oxford Univ. Press, Oxford, pp. 53–71. ↑199

Ganapathy, M. K. 2007. *Robust Mixing*, Electronic Journal of Probability **12**, 262–299. ↑112

Ganapathy, M. and P. Tetali. 2006. *A tight bound for the lamplighter problem*, available at `arXiv:math/0610345v1`. ↑263

Godbole, A. P. and S. G. Papastavridis (eds.) 1994. *Runs and patterns in probability: selected papers*, Mathematics and its Applications, vol. 283, Kluwer Academic Publishers Group, Dordrecht. ↑152

Graham, R. L., D. E. Knuth, and O. Patashnik. 1994. *Concrete mathematics: A foundation for computer science*, second edition, Addison Wesley, Reading, Massachusetts. ↑111, 196

Griffeath, D. 1974/75. *A maximal coupling for Markov chains*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **31**, 95–106. ↑74

Grinstead, C. and L. Snell. 1997. *Introduction to Probability*, 2nd revised edition, American Mathematical Society, Providence. Also available at `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`. ↑35, 60

Häggström, O. 2002. *Finite Markov chains and algorithmic applications*, London Mathematical Society Student Texts, vol. 52, Cambridge University Press, Cambridge. ↑xvi, 60

Häggström, O. 2007. *Problem solving is often a matter of cooking up an appropriate Markov chain*, Scand. J. Statist. **34**, no. 4, 768–780. ↑45

Häggström, O. and J. Jonasson. 1997. *Rates of convergence for lamplighter processes*, Stochastic Process. Appl. **67**, no. 2, 227–249. ↑262, 263

Häggström, O. and K. Nelander. 1998. *Exact sampling from anti-monotone systems*, Statist. Neerlandica **52**, no. 3, 360–380. ↑297

Häggström, O. and K. Nelander. 1999. *On exact simulation of Markov random fields using coupling from the past*, Scand. J. Statist. **26**, no. 3, 395–411. ↑294, 295

Hajek, B. 1988. *Cooling schedules for optimal annealing*, Math. Oper. Res. **13**, no. 2, 311–329. ↑45

Handjani, S. and D. Jungreis. 1996. *Rate of convergence for shuffling cards by transpositions*, J. Theoret. Probab. **9**, no. 4, 983–993. ↑301

Hastings, W. K. 1970. *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika **57**, no. 1, 97 –109. ↑44

Hayes, T. P. and A. Sinclair. 2007. *A general lower bound for mixing of single-site dynamics on graphs*, Ann. Appl. Probab. **17**, no. 3, 931–952, available at `arXiv:math.PR/0507517`. ↑98, 299
Herstein, I. N. 1975. *Topics in algebra*, 2nd ed., John Wiley and Sons, New York. ↑35, 111
Holley, R. and D. Stroock. 1988. *Simulated annealing via Sobolev inequalities*, Comm. Math. Phys. **115**, no. 4, 553–569. ↑45
Hoory, S., N. Linial, and A. Wigderson. 2006. *Expander graphs and their applications*, Bull. Amer. Math. Soc. (N.S.) **43**, no. 4, 439–561 (electronic). ↑188
Horn, R. A. and C. R. Johnson. 1990. *Matrix analysis*, Cambridge University Press, Cambridge. ↑308, 309
Huber, M. 1998. *Exact sampling and approximate counting techniques*, Proceedings of the 30th Annual ACM Symposium on the Theory of Computing, pp. 31–40. ↑294
Ioffe, D. 1995. *Exact large deviation bounds up to T_c for the Ising model in two dimensions*, Probab. Theory Related Fields **102**, no. 3, 313–330. ↑215
Ising, E. 1925. *Beitrag zur theorie der ferromagnetismus*, Zeitschrift Fur Physik **31**, 253–258. ↑215
Jerrum, M. R. 1995. *A very simple algorithm for estimating the number of k-colorings of a low-degree graph*, Random Structures Algorithms **7**, no. 2, 157–165. ↑199
Jerrum, M. 2003. *Counting, sampling and integrating: algorithms and complexity*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel. ↑xvi, 60
Jerrum, M. R. and A. J. Sinclair. 1989. *Approximating the permanent*, SIAM Journal on Computing **18**, 1149–1178. ↑182
Jerrum, M. and A. Sinclair. 1996. *The Markov chain Monte Carlo method: an approach to approximate counting and integration*, Approximation Algorithms for NP-hard Problems. ↑196
Jerrum, M., A. Sinclair, and E. Vigoda. 2004. *A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries*, J. ACM **51**, no. 4, 671–697 (electronic). ↑200
Jerrum, M. R., L. G. Valiant, and V. V. Vazirani. 1986. *Random generation of combinatorial structures from a uniform distribution*, Theoret. Comput. Sci. **43**, no. 2-3, 169–188. ↑200
Johnson, N. L. and S. Kotz. 1977. *Urn models and their application*, John Wiley & Sons, New York-London-Sydney. An approach to modern discrete probability theory; Wiley Series in Probability and Mathematical Statistics. ↑35
Kahn, J. D., N. Linial, N. Nisan, and M. E. Saks. 1989. *On the cover time of random walks on graphs*, J. Theoret. Probab. **2**, no. 1, 121–128. ↑152
Kaĭmanovich, V. A. and A. M. Vershik. 1983. *Random walks on discrete groups: boundary and entropy*, Ann. Probab. **11**, no. 3, 457–490. ↑263
Kakutani, S. 1948. *On equivalence of infinite product measures*, Ann. of Math. (2) **49**, 214–224. ↑274
Kandel, D., Y. Matias, R. Unger, and P. Winkler. 1996. *Shuffling biological sequences*, Discrete Applied Mathematics **71**, 171–185. ↑227
Kantorovich, L. V. 1942. *On the translocation of masses*, C. R. (Doklady) Acad. Sci. URSS (N.S.) **37**, 199–201. ↑191, 199
Kantorovich, L. V. and G. S. Rubinstein. 1958. *On a space of completely additive functions*, Vestnik Leningrad. Univ. **13**, no. 7, 52–59 (Russian, with English summary). ↑200
Karlin, S. and J. McGregor. 1959. *Coincidence properties of birth and death processes*, Pacific J. Math. **9**, 1109–1140. ↑168
Karlin, S. and H. M. Taylor. 1975. *A first course in stochastic processes*, 2nd ed., Academic Press, New York. ↑20
Karlin, S. and H. M. Taylor. 1981. *A second course in stochastic processes*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York. ↑168
Kasteleyn, P. W. 1961. *The statistics of dimers on a lattice I. The number of dimer arrangements on a quadratic lattice*, Physica **27**, no. 12, 1209–1225. ↑320
Keilson, J. 1979. *Markov chain models—rarity and exponentiality*, Applied Mathematical Sciences, vol. 28, Springer-Verlag, New York. ↑168
Kemeny, J. G., J. L. Snell, and A. W. Knapp. 1976. *Denumerable Markov chains*, 2nd ed., Springer-Verlag, New York. With a chapter on Markov random fields, by David Griffeath; Graduate Texts in Mathematics, No. 40. ↑286

Kendall, W. S. and J. Møller. 2000. *Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes*, Adv. in Appl. Probab. **32**, no. 3, 844–865. ↑288

Kenyon, C., E. Mossel, and Y. Peres. 2001. *Glauber dynamics on trees and hyperbolic graphs*, 42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001), IEEE Computer Soc., Los Alamitos, CA, pp. 568–578. ↑207, 214

Knuth, D. 1997. *The art of computer programming*, third edition, Vol. 2: Seminumerical Algorithms, Addison-Wesley, Reading, Massachusetts. ↑319

Kobe, S. 1997. *Ernst Ising—physicist and teacher*, J. Statist. Phys. **88**, no. 3-4, 991–995. ↑215

Kolata, G. January 9, 1990. *In shuffling cards, 7 is winning number*, New York Times, C1. ↑110, 113

Lawler, G. and A. Sokal. 1988. *Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality*, Trans. Amer. Math. Soc. **309**, 557–580. ↑188

Letac, G. 1986. *A contraction principle for certain Markov chains and its applications*, Contemporary Mathematics **50**, 263–273. ↑288

Levin, D. A., M. J. Luczak, and Y. Peres. 2007. *Glauber dynamics for the mean-field Ising model: cut-off, critical power law, and metastability*, available at `arxiv:math.PR/0712.0790`. ↑214, 300

Li, S.-Y. R. 1980. *A martingale approach to the study of occurrence of sequence patterns in repeated experiments*, Ann. Probab. **8**, no. 6, 1171–1176. ↑245

Lindvall, T. 2002. *Lectures on the coupling method*, Dover, Mineola, New York. ↑74

Littlewood, J. E. 1948. *Large Numbers*, Mathematical Gazette **32**, no. 300. ↑99

Lovász, L. 1993. *Random walks on graphs: a survey*, Combinatorics, Paul Erdős is Eighty, pp. 1–46. ↑60, 143

Lovász, L. and R. Kannan. 1999. *Faster mixing via average conductance*, Annual ACM Symposium on Theory of Computing (Atlanta, GA, 1999), ACM, New York, pp. 282–287 (electronic). ↑246

Lovász, L. and P. Winkler. 1993. *On the last new vertex visited by a random walk*, J. Graph Theory **17**, 593–596. ↑84

Lovász, L. and P. Winkler. 1995a. *Exact mixing in an unknown Markov chain*, Electronic Journal of Combinatorics **2**. Paper #R15. ↑296

Lovász, L. and P. Winkler. 1995b. *Efficient stopping rules for Markov chains*, Proc. 27th ACM Symp. on the Theory of Computing, pp. 76–82. ↑85

Lovász, L. and P. Winkler. 1998. *Mixing times*, Microsurveys in discrete probability (Princeton, NJ, 1997), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 41, Amer. Math. Soc., Providence, RI, pp. 85–133. ↑60, 83, 85, 264

Loynes, R. M. 1962. *The stability of a queue with non-independent inter-arrival and service times*, Proceedings of the Cambridge Philosophical Society **58**, 497–520. ↑288

Lubetzky, E. and A. Sly. 2008. *Cutoff phenomena for random walks on random regular graphs*. in preparation. ↑255

Lubotzky, A. 1994. *Discrete groups, expanding graphs and invariant measures*, Progress in Mathematics, vol. 125, Birkhäuser Verlag, Basel. With an appendix by Jonathan D. Rogawski. ↑188

Luby, M., D. Randall, and A. Sinclair. 1995. *Markov chain algorithms for planar lattice structures*, Proceedings of the 36th IEEE Symposium on Foundations of Computing, pp. 150–159. ↑74, 320

Luby, M., D. Randall, and A. Sinclair. 2001. *Markov chain algorithms for planar lattice structures*, SIAM J. Comput. **31**, 167–192. ↑74

Luby, M. and E. Vigoda. 1995. *Approximately counting up to four*, Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, pp. 150–159. Extended abstract. ↑74

Luby, M. and E. Vigoda. 1999. *Fast convergence of the Glauber dynamics for sampling independent sets*, Random Structures and Algorithms **15**, no. 3-4, 229–241. ↑74

Lyons, R. and Y. Peres. 2008. *Probability on Trees and Networks*, in progress. Manuscript available at `http://php.indiana.edu/~rdlyons/prbtree/prbtree.html`. ↑125, 286

Lyons, T. 1983. *A simple criterion for transience of a reversible Markov chain*, Ann. Probab. **11**, no. 2, 393–402. ↑278, 286

Madras, N. and D. Randall. 1996. *Factoring graphs to bound mixing rates*, Proceedings of the 37th IEEE Symposium on Foundations of Computing, pp. 194–203. ↑188

Madras, N. and G. Slade. 1993. *The self-avoiding walk*, Birkhäuser, Boston. ↑320

Mann, B. 1994. *How many times should you shuffle a deck of cards?*, UMAP J. **15**, no. 4, 303–332. ↑113

Markov, A. A. 1906. *Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga*, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya **15**, 135–156. ↑20

Martinelli, F. 1994. *On the two-dimensional dynamical Ising model in the phase coexistence region*, J. Statist. Phys. **76**, no. 5-6, 1179–1246. ↑299

Martinelli, F. 1999. *Lectures on Glauber dynamics for discrete spin models*, Lectures on probability theory and statistics (Saint-Flour, 1997), Lecture Notes in Math., vol. 1717, Springer, Berlin, pp. 93–191. ↑211, 215, 300

Martinelli, F. and E. Olivieri. 1994. *Approach to equilibrium of Glauber dynamics in the one phase region. I. The attractive case*, Comm. Math. Phys. **161**, no. 3, 447–486. ↑215

Martinelli, F., E. Olivieri, and R. H. Schonmann. 1994. *For 2-D lattice spin systems weak mixing implies strong mixing*, Comm. Math. Phys. **165**, no. 1, 33–47. ↑215

Martinelli, F., A. Sinclair, and D. Weitz. 2004. *Glauber dynamics on trees: boundary conditions and mixing time*, Comm. Math. Phys. **250**, no. 2, 301–334. ↑214

Matthews, P. 1988a. *Covering problems for Markov chains*, Ann. Probab. **16**, 1215–1228. ↑144, 152

Matthews, P. 1988b. *A strong uniform time for random transpositions*, J. Theoret. Probab. **1**, no. 4, 411–423. ↑112

Matthews, P. 1989. *Some sample path properties of a random walk on the cube*, J. Theoret. Probab. **2**, no. 1, 129–146. ↑152

McKay, B. D. and C. E. Praeger. 1996. *Vertex-transitive graphs that are not Cayley graphs. II*, J. Graph Theory **22**, no. 4, 321–334. ↑29

Metropolis, N., A. W. Rosenbluth, A. H. Teller, and E. Teller. 1953. *Equation of state calculations by fast computing machines*, J. Chem. Phys. **21**, 1087–1092. ↑44

Mihail, M. 1989. *Conductance and convergence of Markov chains - A combinatorial treatment of expanders*, Proceedings of the 30th Annual Conference on Foundations of Computer Science, 1989, pp. 526–531. ↑98

Mironov, I. 2002. *(Not so) random shuffles of RC4*, Advances in cryptology—CRYPTO 2002, Lecture Notes in Comput. Sci., vol. 2442, Springer, Berlin, pp. 304–319. ↑112

Montenegro, R. and P. Tetali. 2006. *Mathematical aspects of mixing times in Markov chains*, Vol. 1. ↑xvi, 60

Móri, T. F. 1987. *On the expectation of the maximum waiting time*, Ann. Univ. Sci. Budapest. Sect. Comput. **7**, 111–115 (1988). ↑152

Morris, B. 2008. *Improved mixing time bounds for the Thorp shuffle and L-reversals*, available at `arXiv:math.PR/0802.0339`. ↑227

Morris, B. and Y. Peres. 2005. *Evolving sets, mixing and heat kernel bounds*, Probab. Theory Related Fields **133**, no. 2, 245–266. ↑235, 246

Mossel, E., Y. Peres, and A. Sinclair. 2004. *Shuffling by semi-random transpositions*, Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04) October 17 - 19, 2004, Rome, Italy, pp. 572–581. ↑112, 187, 300

Nacu, Ş. 2003. *Glauber dynamics on the cycle is monotone*, Probab. Theory Related Fields **127**, no. 2, 177–185. ↑299

Nash-Williams, C. St. J. A. 1959. *Random walks and electric currents in networks*, Proc. Cambridge Philos. Soc. **55**, 181–194. ↑125

von Neumann, J. 1951. *Various techniques used in connection with random digits*, National Bureau of Standards Applied Mathematics Series **12**, 36–38. ↑312, 325

Norris, J. R. 1998. *Markov chains*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 2, Cambridge University Press, Cambridge. Reprint of 1997 original. ↑xvi, 286

Pak, I. 1997. *Random Walks on Groups : Strong Uniform Time Approach*, Ph.D. thesis, Harvard University. ↑85

Pemantle, R. 2007. *A survey of random processes with reinforcement*, Probab. Surv. **4**, 1–79 (electronic). ↑35

Peres, Y. 1992. *Iterating von Neumann's procedure for extracting random bits*, Ann. Stat. **20**, no. 1, 590–597. ↑313, 325

Peres, Y. 1999. *Probability on trees: an introductory climb*, Lectures on Probability Theory and Statistics, Ecole d'Ete de Probabilites de Saint-Flour XXVII - 1997, pp. 193–280. ↑125

Peres, Y. 2002. *Brownian intersections, cover times and thick points via trees*, (Beijing, 2002), Higher Ed. Press, Beijing, pp. 73–78. ↑152

Peres, Y. and D. Revelle. 2004. *Mixing times for random walks on finite lamplighter groups*, Electron. J. Probab. **9**, no. 26, 825–845 (electronic). ↑263

Pinsker, M. S. 1973. *On the complexity of a concentrator*, Proc. 7th Int. Teletraffic Conf., Stockholm, Sweden, pp. 318/1–318/4. ↑186, 188

Pisztora, A. 1996. *Surface order large deviations for Ising, Potts and percolation models*, Probab. Theory and Related Fields **104**, no. 4, 427–466. ↑215

Pitman, J. W. 1974. *Uniform rates of convergence for Markov chain transition probabilities*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **29**, 193–227. ↑74

Pitman, J. W. 1976. *On coupling of Markov chains*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **35**, no. 4, 315–322. ↑74

Pólya, G. 1931. *Sur quelques points de la théorie des probabilités*, Ann. Inst. H. Poincaré **1**, 117–161. ↑35

Propp, J. and D. Wilson. 1996. *Exact sampling with coupled Markov chains and applications to statistical mechanics*, Random Structure and Algorithms **9**, 223–252. ↑287, 288, 293

Propp, J. and D. Wilson. 1998. *How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph*, Journal of Algorithms (SODA '96 special issue) **27**, no. 2, 170–217. ↑296, 297

Quastel, J. 1992. *Diffusion of color in the simple exclusion process*, Comm. Pure Appl. Math. **45**, no. 6, 623–679. ↑182, 188

Randall, D. 2006. *Slow mixing of Glauber dynamics via topological obstructions*, SODA (2006). Available at `http://www.math.gatech.edu/~randall/reprints.html`. ↑211, 213

Randall, D. and A. Sinclair. 2000. *Self-testing algorithms for self-avoiding walks*, Journal of Mathematical Physics **41**, no. 3, 1570–1584. ↑320, 322

Randall, D. and P. Tetali. 2000. *Analyzing Glauber dynamics by comparison of Markov chains*, J. Math. Phys. **41**, no. 3, 1598–1615. Probabilistic techniques in equilibrium and nonequilibrium statistical physics. ↑188

Riordan, J. 1944. *Three-line Latin rectangles*, Amer. Math. Monthly **51**, 450–452. ↑187

Röllin, A. 2007. *Translated Poisson approximation using exchangeable pair couplings*, Ann. Appl. Probab. **17**, no. 5-6, 1596–1614, available at `arxiv:math.PR/0607781`. ↑274

Salas, J. and A. D. Sokal. 1997. *Absence of phase transition for antiferromagnetic Potts models via the Dobrushin uniqueness theorem*, J. Statist. Phys. **86**, no. 3-4, 551–579. ↑199

Saloff-Coste, L. 1997. *Lectures on finite Markov chains*, Lectures on Probability Theory and Statistics, Ecole d'Ete de Probabilites de Saint-Flour XXVI - 1996, pp. 301–413. ↑60, 141

Sarnak, P. 2004. *What is... an expander?*, Notices Amer. Math. Soc. **51**, no. 7, 762–763. ↑188

Scarabotti, F. and F. Tolli. 2008. *Harmonic analysis of finite lamplighter random walks*, J. Dyn. Control Syst. **14**, no. 2, 251–282, available at `arXiv:math.PR/0701603`. ↑263

Schonmann, R. H. 1987. *Second order large deviation estimates for ferromagnetic systems in the phase coexistence region*, Comm. Math. Phys. **112**, no. 3, 409–422. ↑211, 214

Seneta, E. 2006. *Non-negative matrices and Markov chains*, Springer Series in Statistics, Springer, New York. Revised reprint of the second (1981) edition [Springer-Verlag, New York]. ↑20, 60

Simon, B. 1993. *The statistical mechanics of lattice gases. Vol. I*, Princeton Series in Physics, Princeton University Press, Princeton, NJ. ↑215

Sinclair, A. 1992. *Improved bounds for mixing rates of Markov chains and multicommodity flow*, Combin. Probab. Comput. **1**, no. 4, 351–370. ↑188

Sinclair, A. 1993. *Algorithms for random generation and counting*, Progress in Theoretical Computer Science, Birkhäuser Boston Inc., Boston, MA. A Markov chain approach. ↑60, 200

Sinclair, A. and M. Jerrum. 1989. *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. and Comput. **82**, no. 1, 93–133. ↑188, 212

Snell, J. L. 1997. *A conversation with Joe Doob*, Statist. Sci. **12**, no. 4, 301–311. ↑303

Soardi, P. M. 1994. *Potential theory on infinite networks*, Lecture Notes in Mathematics, vol. 1590, Springer-Verlag, Berlin. ↑286

Spielman, D. A. and S.-H. Teng. 1996. *Spectral partitioning works: planar graphs and finite element meshes*, 37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996), IEEE Comput. Soc. Press, Los Alamitos, CA. ↑168

Spitzer, F. 1976. *Principles of random walks*, 2nd ed., Springer-Verlag, New York. Graduate Texts in Mathematics, Vol. 34. ↑273

Stanley, R. P. 1986. *Enumerative combinatorics*, Vol. 1, Wadsworth & Brooks/Cole, Belmont, California. ↑196

Stanley, R. P. 1999. *Enumerative Combinatorics*, Vol. 2, Cambridge University Press. ↑35

Stanley, R. P. 2008. *Catalan Addendum*. Available at `http://www-math.mit.edu/~rstan/ec/catadd.pdf`. ↑35

Stroock, D. W. and B. Zegarliński. 1992. *The equivalence of the logarithmic Sobolev inequality and the Dobrushin-Shlosman mixing condition*, Comm. Math. Phys. **144**, no. 2, 303–323. ↑215

Sugimine, N. 2002. *A lower bound on the spectral gap of the 3-dimensional stochastic Ising models*, J. Math. Kyoto Univ. **42**, no. 4, 751–788 (2003). ↑299

Thomas, L. E. 1989. *Bound on the mass gap for finite volume stochastic Ising models at low temperature*, Comm. Math. Phys. **126**, no. 1, 1–11. ↑211, 212, 214

Thorisson, H. 1988. *Backward limits*, Annals of Probability **16**, 914–924. ↑288

Thorisson, H. 2000. *Coupling, stationarity, and regeneration*, Probability and its Applications (New York), Springer-Verlag, New York. ↑74, 286

Thorp, E. O. 1965. *Elementary Problem E1763*, Amer. Math. Monthly **72**, no. 2, 183. ↑111, 112

Thurston, W. P. 1990. *Conway's tiling groups*, Amer. Math. Monthly **97**, no. 8, 757–773. ↑9

Uyemura-Reyes, J. C. 2002. *Random Walk, Semidirect Products, and Card Shuffling*, Ph.D. thesis, Stanford University. ↑300

Vasershtein, L. N. 1969. *Markov processes over denumerable products of spaces describing large system of automata*, Problemy Peredači Informacii **5**, no. 3, 64–72 (Russian); English transl.,. 1969, Problems of Information Transmission **5**, no. 3, 47–52. ↑199

Vershik, A. M. 2004. *The Kantorovich metric: the initial history and little-known applications*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **312**, no. Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 11, 69–85, 311 (Russian, with English and Russian summaries); English transl.,. 2004, J. Math. Sci. (N. Y.) **133**, no. 4, 1410–1417, available at `arxiv:math.FA/0503035`. ↑199

Vigoda, E. 2000. *Improved bounds for sampling colorings*, J. Math. Phys. **41**, no. 3, 1555–1569. ↑199

Vigoda, E. 2001. *A note on the Glauber dynamics for sampling independent sets*, Electron. J. Combin. **8**, no. 1, Research Paper 8, 8 pp. (electronic). ↑74, 295

Villani, C. 2003. *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI. ↑199

Wilf, H. S. 1989. *The editor's corner: The white screen problem*, Amer. Math. Monthly **96**, 704–707. ↑152

Williams, D. 1991. *Probability with martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge. ↑245

Wilson, D. B. 2000a. *Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP)* (N. Madras, ed.), Fields Institute Communications, vol. 26, American Mathematical Society. ↑297

Wilson, D. B. 2000b. *How to couple from the past using a read-once source of randomness*, Random Structures and Algorithms **16**, 85–113. ↑288, 292

Wilson, D. B. 2003. *Mixing time of the Rudvalis shuffle*, Electron. Comm. Probab. **8**, 77–85 (electronic). ↑187

Wilson, D. B. 2004a. *Mixing times of Lozenge tiling and card shuffling Markov chains*, Ann. Appl. Probab. **14**, no. 1, 274–325. ↑187, 218, 227, 320

Wilson, D. B. 2004b. *Perfectly Random Sampling with Markov Chains*. Available at `http://research.microsoft.com/~dbwilson/exact/`. ↑297

Woess, W. 2000. *Random walks on infinite graphs and groups*, Cambridge Tracts in Mathematics, vol. 138, Cambridge University Press, Cambridge. ↑286

Zuckerman, D. 1989. *Covering times of random walks on bounded degree trees and other graphs*, J. Theoret. Probab. **2**, no. 1, 147–157. ↑152

Zuckerman, D. 1992. *A technique for lower bounding the cover time*, SIAM J. Discrete Math. **5**, 81–87. ↑152

van Zuylen, A. and F. Schalekamp. 2004. *The Achilles' heel of the GSR shuffle. A note on new age solitaire*, Probab. Engrg. Inform. Sci. **18**, no. 3, 315–328. ↑113

Notation Index

The symbol := means *defined as.*

The set $\{\dots,-1,0,1,\dots\}$ of integers is denoted $\mathbb{Z}$ and the set of real numbers is denoted $\mathbb{R}$.

For sequences (a_n) and (b_n), the notation $a_n = O(b_n)$ means that for some $c > 0$ we have $a_n/b_n \leq c$ for all n, while $a_n = o(b_n)$ means that $\lim_{n\to\infty} a_n/b_n = 0$, and $a_n \asymp b_n$ means both $a_n = O(b_n)$ and $b_n = O(a_n)$ are true.

Index

Italics indicate that the reference is to an exercise.

本书是对 Markov 链理论的现代处理方法的导引，该方法的主要目标是确定一个 Markov 链收敛到作为态空间体积和几何之函数的平稳分布的收敛速率。作者发展了估计收敛时间的关键工具，包括耦合、强平稳时间以及谱方法；一有可能，便强调概率论式的方法。本书包括了许多例题并对统计力学的中心模型给出了简短介绍；还讲述了网络上的随机游动，包括击中和掩盖时间，以及对洗牌的各种方法的分析。至于预备知识，作者假定了对概率论的适度了解以及大学水平的线性代数。本书打算将这个活跃的研究领域的激情带给大范围的受众。

美国数学会经典影印系列
AMS
AMERICAN MATHEMATICAL SOCIETY

《Markov 链与混合时间》是一本神奇的书，处理得既平易亲切又很深刻。它悄然地引进了概率论式的技术，使得一个门外汉也能跟得上。同时，它是第一本包含了 Markov 链的几何理论的书，并有许多甚至对专家来说也很新的内容。它肯定是我要用来教课的书。我推荐它给所有的新手，它是一个了不起的成就。

—*Persi Diaconis, Mary V. Sunseri* 统计学和数学教授，*Stanford Univrsity*

在这本书中，作者迅速地将一个完全准备好了的大学生带到了研究的前沿。简短并伴以清晰的逻辑关联的章节让读者可以以多重的方式来使用这本书。

—*CHOICE Magazine*

作者通篇都慷慨地给出了启发理论和诠释思想的例题。我期望这本杰出的书能被广泛地用作世界各地的研究生课程，并成为一本标准的参考书。

—*Mathematical Reviews*

ISBN 978-7-04-046994-3



定价 169.00 元